HANDBOOK OF TELECOMMUNICATIONS ECONOMICS

Volume 1

Martin E. Cave

HANDBOOK OF TELECOMMUNICATIONS ECONOMICS
VOLUME 1

This Page Intentionally Left Blank

HANDBOOK OF TELECOMMUNICATIONS ECONOMICS

VOLUME 1

STRUCTURE, REGULATION AND COMPETITION

Edited by

MARTIN E. CAVE
University of Warwick

SUMIT K. MAJUMDAR
Imperial College of Science, Technology and Medicine
University of London

and

INGO VOGELSANG
Boston University

2002
ELSEVIER

AMSTERDAM • BOSTON • LONDON • NEW YORK • OXFORD • PARIS
SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2002
Reprinted 2005, 2006

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN–13: 978-0-444-50389-3
ISBN–10: 0-444-50389-7

ISSN- 1569-4054

Printed and bound in *The Netherlands*

06 07 08 09 10 10 9 8 7 6 5 4 3

LIST OF CONTRIBUTORS

Dr Mark Armstrong	Economics Group Nuffield College Oxford University Oxford, UK
Professor Gerald W. Brock	Graduate Telecommunication Program The George Washington University Washington, USA
Professor Martin E. Cave	Centre for Management under Regulation University of Warwick Coventry, UK
Professor Peter Cramton	Department of Economics University of Maryland College Park, USA
Professor Melvyn A. Fuss	Department of Economics University of Toronto Toronto, Canada
Professor Jerry Hausman	Department of Economics Massachusetts Institute of Technology Cambridge, USA
Professor David L. Kaserman	College of Business Auburn University Auburn, USA
Professor Stanley J. Liebowitz	School of Management University of Texas at Dallas Richardson, USA
Professor Sumit K. Majumdar	The Management School Imperial College of Science, Technology and Medicine University of London London, UK
Professor John W. Mayo	McDonough School of Business Georgetown University Washington, USA

Professor Stephen E. Margolis — College of Management, North Carolina State University, Raleigh, USA

Professor Eli M. Noam — School of Business, Columbia University, New York, USA

Professor Michael H. Riordan — Department of Economics, Graduate School of Business, New York, USA

Professor Daniel F. Spulber — Department of Management and Strategy, J.L. Kellogg Graduate School of Management, Northwestern University, Evanston, USA

Dr. William W. Sharkey — Federal Communications Commission, Washington DC, USA

Professor David E. M. Sappington — Lanzilloti-McKethan Eminent Scholar, Department of Economics, University of Florida, Gainesville, USA

Professor Lester D. Taylor — Department of Economics, University of Arizona, USA

Professor Ingo Vogelsang — Department of Economics, Boston University, Boston, USA

Professor Glen A. Woroch — Department of Economics, University of California, Berkeley, USA

Professor Leonard Waverman — London Business School, London, UK

CONTENTS OF VOLUME 1 OF THE HANDBOOK

SECTION I – STRUCTURE

SECTION II – REGULATION

SECTION III – COMPETITION

Chapter 1

STRUCTURE, REGULATION AND COMPETITION IN THE TELECOMMUNICATIONS INDUSTRY*

MARTIN E. CAVE

University of Warwick

SUMIT K. MAJUMDAR

Imperial College of Science, Technology and Medicine, University of London

INGO VOGELSANG

Boston University

Contents

* This Handbook would not have seen the light of day but for the untiring efforts of our administrative colleagues Ms. Dannie Carr at *Imperial College* and Ms. Marion Cherrie at *Brunel University*. Their professional approach and good cheer has been an extraordinary input towards the completion of the project.

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

The last two decades have seen exceptionally fast rates of change in every aspect of the telecommunications industry. These include technology changes – especially the effects of digitalization; introduction of new products including internet-based services, and convergence associated with the coming together of the broadcasting, information technology and telecommunications industries. Simultaneously, market structure has changed through the replacement of the former monopolistic, vertically integrated telephone companies by a variety of competing firms. These developments have been accompanied by major legislative and regulatory developments, including the passage in the United States of the 1996 Telecommunications Act and the introduction of a large number of new laws and regulations in Europe and elsewhere. The same changes have seen a massive expansion of independent regulatory agencies. Their major role has been to determine the terms of competitive interactions between the existing incumbent operators and new entrants.

This volume provides an economic review of and commentary on these changes in the landscape of what is arguably a very major industry. It attempts to provide a comprehensive account of the major applications of economic analysis to our field of interest, which is essentially two-way communications between parties that are not in physical contact with each other by means of fixed line telephony, mobile telephony, satellite telephony, cable TV telephony and Internet telephony. To a significant extent, voice communication is the process of interest. We propose in a second volume to focus our attention on the economic analysis of data communications, including in particular the development and use of the Internet, and on a number of other strategic, institutional and organizational issues of the incredibly fast-moving contemporary telecommunications landscape.

The three principal sections of this volume contain fourteen chapters written by experts in the relevant areas. Each chapter is intended to be self-contained, and can be read independently. Together, however, the chapters provide a survey of the bulk of recent work in the economics of the telecommunications sector. The reader pursuing a particular subject, for example the application of engineering cost models, or the relation between wholesale and retail price regulation, is likely to want to consult several of them.

Authorship of the chapters by different individuals carries the advantage that readers can gain access to their special expertise on the subject in question. It also exposes differences among authors that may exist on key current policy questions. For example, some authors, who have greater confidence in the potentially benevolent effects of regulation than others, would like to see the process of deregulation taking place more quickly. This diversity of opinion is an advantage, because it reflects genuine disagreement among economists and policy makers on matters which are still of major topical concern. Kuhn (1962) has described the early stages of the evolution of a field as

consisting of phenomena reporting in the absence of a unifying theoretical framework. Thus, a plurality of perspectives may be observed being brought to bear on issues. Even in the presence of theories that may be seemingly logical, the empirical phenomena noted might be at odds with the assumptions that underlie the framework. The existence of multiple views enables us, as editors, in this introductory chapter, to highlight some of these differences without necessarily attempting to reconcile them. It also enables us to articulate the key conceptual issues that we think ought to form the core of the discipline of telecommunications economics that is evolving.

Although we have sought to be comprehensive, particularly in explaining concepts that underlie policy making and in reviewing the literature, the single most important bias of the current volume is its focus on highly developed countries, the U.S. in particular. The main reasons are that the most important work has been on American telecommunications problems and that most of our authors are U.S. based. Critical policy experiments were carried out in the U.S., and the extant literature covers U.S. experience in the main. Nevertheless, general consequences from U.S. policies are lessons drawn by the authors and in this introduction.

Two features of the U.S. telecommunications sector are of particular importance – the ownership structure and the degree of vertical integration. Most countries in the world have a dominant telecommunications carrier that is or was state owned. In contrast, the U.S. Bell system was always in private hands[1]. In spite of this difference, the resulting policy issues and outcomes are strikingly similar across countries. The U.S. telecommunication sector has been less vertically integrated than elsewhere, since AT&T was split up in 1984. This vertical separation has provided an interesting policy experiment and has justified a separate chapter on long-distance competition.

Our second bias is the omission of macro-economic consequences of telecommunications. The design of regulatory institutions was planned as a chapter but could not be completed on time. However, this area is partially covered in several chapters, notably in the chapters dealing with the historical overview, in interconnection practice, mobile telephony and local network competition. We also take it up in this introduction.

The contents of the *Handbook* are set out in three main sections. The first section deals with the economic characteristics of the sector. These characteristics define the structure of the industry, and we discuss how different aspects of structure can interact to influence industry development. The next section deals with regulation and the third section deals with competition. In theory, the section on competition should precede that on regulation. In practice, regulatory issues

[1] This particular volume of the *Handbook* does not cover the topic of privatization. Nevertheless, given the salience of the topic, some comments are in order and these comments are included at various points in this introductory chapter.

have dominated the policy debates for quite some time. While issues related to competition are critical, they are of more recent origin. Of course, the regulatory issues and competition issues continuously interact with each other, and each set of issues influences the other. The industry evolves through the interaction of the two processes. We invite the reader to keep this in mind while reading the chapters within the three sections.

2. Economic characteristics

The objective of the first section of the *Handbook* is to set out some history as well as the crucial economic and structural characteristics of the telecommunications sector, so that policymaking and firms' decisions are guided by the appropriate principles. In designing the ordering of the chapters in this section, the basic organizing principle followed is that structural issues of demand and supply are sequentially looked at in detail. These demand and supply issues then underpin the theoretical discussions related to all of the other topics dealt with in the volume.

Following the "Historical Overview" of the telecommunications industry in the United States by Gerald Brock (Brock), two chapters deal with the issues of demand. In their chapter, Stanley Liebowitz and Stephen Margolis (Liebowitz and Margolis) deal with "Network Effects" that gain particular importance in an industry setting where interconnectivity is a unique characteristic. These authors take a relatively macro-level look at the industry-level consumption characteristics and provide the backdrop for the chapter by Lester Taylor (Taylor), who looks at micro-level "Customer Demand Analysis" issues in the sector.

Following the two demand-side chapters, Melvyn Fuss and Leonard Waverman (Fuss and Waverman) analyze "Econometric Cost Functions," and William Sharkey (Sharkey) provides a "Representation of Technology and Production." Both these chapters deal with the supply-side of the telecommunications industry equation. The chapter by Fuss and Waverman is pitched at a relatively aggregate level of analysis, while the one by Sharkey is written intrinsically at a more micro-level of analysis.

2.1. Historical overview

Gerald Brock's "Historical Overview" is a broad review of the evolution of the telecommunications industry in the United States. It is an entirely country-centric chapter. Yet, there are crucial lessons to be learnt from the United States' experience by other countries around the world for several reasons. First, of course, at a basic technological level the telephone was invented in the United States. Nevertheless, the industry evolution patterns and public policy evolution patterns have much salience for other countries. In the United States, after the

expiration of the basic patents held by the Bell system towards the end of the nineteenth century a number of new telephone companies entered the industry. These companies expanded so rapidly that prices fell sharply. As a result of entry a considerable enhancement of infrastructure took place so that not only were large cities such as Boston and New York supplied with telephone lines, but deep rural portions of the country had substantial penetration as well. And, this degree of penetration had taken place at the turn of the last century! It is thus well worth taking into account the lessons from history[2].

In the light of this decline in prices and increase in investment in infrastructure that was observed almost a century ago, when monopoly power dissolved, the deep and fundamental question remains, what should be public policy towards reducing the monopoly of contemporary operators so that consumer welfare is enhanced? The evidence from the early institutional history of the United States that Brock presents shows that when the institutional context changed away from monopoly to a competitive market then entry of new firms was rapid and prices fell, at least initially. Whether such benefits would have existed or not if entry was not permitted is a counter-factual issue. However, from Brock's chapter it is clear that the salient contemporary issues of access and interconnection raised their heads at the turn of the century.

Theodore Vail, as president of American Telephone and Telegraph Company (AT&T), led a process of industry consolidation and the creation of a telecommunications conglomerate. He bought up smaller operators after realising that interconnection was technologically necessary so that customers belonging to a multitude of phone companies could talk to each other and not just to subscribers of one particular network. Implicit in his plan was the need to exploit economies of scale and scope[3]. However, as Chandler (1977), in his seminal work on the theory of the firm, has noted, such economies came in a large measure from the administrative efficiencies in handling through traffic that interconnectivity would permit. Such administrative efficiencies also generated system-wide resources to be invested in implementing an expansion policy for the Bell system as a whole. The planning for such system-wide resource deployment and expansion could be done from the head office of the system. Thus, the sources of economies were organisational and not just technological[4].

[2] The historical events described in this chapter are also treated in the chapters written by Eli Noam and Glenn Woroch for this *Handbook*.

[3] However, the elucidation and empirical examination of scale and scope economies only began to take place much later in the 1960s.

[4] In fact, it was suggested that the operating costs per subscriber increased with the number of subscribers and that there were diminishing returns to scale in the construction and operation of telephone exchanges. Thus, cost efficiencies, if any, would come through administrative efficiencies that would be enjoyed by one large firm operating in the sector relative to two small firms operating in the sector. The administrative efficiencies of one large firm would outweigh the combined operating efficiencies enjoyed by two small firms. Thus, the issue of subadditivity and the recognition that economies of scope were organizational in nature were important conceptual constructs a century ago.

Access by companies to each other's networks might have solved the technological interconnection problem. Whether, however, Theodore Vail was justified, given his search for administrative economies and the desire to establish a central strategic planning apparatus, in creating a monopoly is moot. Between the expiration of the Bell system patents in 1893 and 1894 and 1907 independent telephone companies operated 51 percent of the telephone lines in the U.S. Between 1907 and 1912 the Bell system started absorbing independent companies into its fold and the market share of independent companies dropped. The presence of competition between telephone companies for franchises and between franchises for subscribers clearly stimulated the extension of telephone service, the reduction of telephone rates, and technological improvements in telephony. The process of consolidation that was started by AT&T at the turn of the century halted these positive outcomes.

Today it is recognised that a competitive market structure is the viable option if markets expand and economies of scale and scope are exhausted at small operator size. Other writers on the history of the U.S. telecommunications industry also conclude that clearly competition was better than unregulated monopoly. But whether the continuation of competition in the U.S. would have been better than the presence of a regulated monopoly in continuing to yield such superior outcomes will remain unknown[5]. In Brock's chapter, the United States experience moves from one of an early competitive market structure to that of a regulated monopoly, which then persisted for almost three-quarters of a century till the break up of the Bell system in 1984. Starting in the 1950s, a process of gradual introduction of competition in the customer premises equipment, long distance and local segments has taken place. Such moves are being copied all over the world.

Another crucial aspect of the United States industry is the role of the private sector. AT&T was a regulated privately owned firm. As Brock observes, AT&T chose to accept regulation and use it as a substitute for market forces. Yet, the regulation of the private company was instrumental to the enormous infrastructure development that took place in the United States. What were the aspects of the regulatory regimes in place that led to the development of the telecommunications infrastructure to the degree that it did in the United States? This question becomes important as nations around the world are struggling with questions of appropriate institutional design for their telecommunications sector.

Allied to these changes is the design of regulatory instruments and organisations. From the United States case it is apparent that public ownership, *per se*, is not a prerequisite for effective infrastructure development. This is a theme, once novel, now increasingly taken as axiomatic around the world. What has perhaps

[5] Nevertheless, a disquiet that a regulated monopoly was impeding technical progress and creating X-inefficiencies was implicit in fuelling the process of litigation and institutional changes that eventually led to the re-structuring of AT&T in 1984 (Shepherd, 1983).

led to the development of the U.S. telecommunications infrastructure is the decentralisation of regulatory activities, to the extent possible, and effective and stable implementation of regulatory policies and rules under rate-of-return regulation. As far as the former aspect is concerned, there was, and still is, a regulatory body in each of the states as well as the Federal Communications Commission at the federal level. With respect to implementation of regulations, the co-ordinated application of the separations rules is one of the factors that has led to the widespread diffusion of the basic telephone service, which most recently was at 95 percent. Thus, the design and implementation of rules do matter very clearly. In the next section of the *Handbook*, several chapters deal with contemporary regulatory design issues.

Nevertheless, as Brock also observes, in the face of technical change, entrenched monopolies can use the regulatory process to play entry deterrence games. AT&T played these games so that the Federal Communications Commission (FCC) prohibited other firms from developing specialised video transmission facilities. These games may be currently occurring as telecommunications markets are being opened to technology-driven competition. Brock's chapter implicitly brings out, from a historical perspective, some of the organisational and processual dimensions involved in the design and implementation of regulatory institutions.

A second issue emerging from Brock's chapter is the role that should be accorded to the FCC's Computer Inquiry II decision (Computer II). This decision has had far-reaching impact, and is an interesting institutional decision to study. First, the Computer II decision explicitly impacted on the boundaries of the telecommunications firm. By separating customer premises equipment (CPE) provision from telecommunications services, the Computer II decision clearly stipulated the boundaries of telephony service provision[6]. By making CPE provision unregulated, the Computer II decision led to the very large growth of the CPE segment of the industry. Thus, freeing up a segment to competition has had enormous impact on new firm entry and the evolution of an industry segment.

Second, the Computer II decision has played a pivotal role in providing the basis for regulating access to the Internet in the United States. The Computer II decision created a right for customers to provide services within their own premises and to connect those services to the public switched telephone network (PSTN). The principles of Computer II are currently applied to regulation of the Internet through recognition that intelligence is generated and retained by the Internet service providers (ISPs) at the edges of the network. This happens with CPE within a telephone network. ISPs gain access to the PSTN without having

[6] Of course, empirical evaluations of boundary separations are far from ever being perfect. The need for empirical analysis to ascertain where boundaries of different sectors might lie, howsoever fuzzy the measurement, remains crucial.

to pay the often-high access charges that other telephone companies have to pay the incumbents for access to their PSTNs. This implication of Computer II, permitting ISPs almost free access to the PSTN, has resulted in the most extraordinary diffusion of the Internet in the United States. This had consequences for global consumers as well, since a very large proportion of the Internet content is generated in the United States. The availability of the large United States customer base has generated a large number of content-developing enterprises whose output is circulated globally. The Computer II decision has had global consequences.

2.2. Network effects

The chapter by Stan Liebowitz and Steve Margolis (Liebowitz and Margolis) on "Network Effects" deals with the most fundamental economic characteristic of a communications system. The importance of connectivity and the value placed by consumers on that feature is well recognised. In their chapter, Liebowitz and Margolis deal broadly with issues of classification of network effects, features of network effects, and the somewhat limited empirical literature[7].

Liebowitz and Margolis highlight the importance of the role of network size and network choice. From their articulation of these issues one can draw important implications to current regulatory controversies. First, the ability to capture and internalize externalities via ownership or market position is an inducement for the development of large networks. The network effects generated in such networks provide profit opportunities for the owner. In this context there are close intellectual parallels between the network externalities and the natural monopoly arguments: each sustains the position of the regulated incumbent monopolist. Certainly, both consumption and production can exhibit increasing returns to network size or scope. Such conditions would militate against some forms of unbundling if the efficiencies enjoyed by the large network operators were to be lost as a result. On the other hand, with mixed bundling, specific elements of the local loop are leased out. Such lease arrangements exploit the spare capacities available on local loop networks. In such a case, any loss of network efficiencies may well be small and may be offset by the leasing revenues received. In view of the ramifications of public policy toward networks, precise measurement and quantification of scale and scope economies should be a high priority.

Second, several small networks might interconnect. This has happened with the Internet. In that case, the consumption scale economies may be insufficiently small for any network operator to make meaningful gains because of limited

[7] Starting from early work (Rohlfs, 1974), authors in the telecommunications field have shown a great deal of interest in consumption externalities on the demand side. Of more recent origin is the notion of production externalities, which are the supply side effects arising from connectivity between networks.

network size[8]. Concomitantly, congestion effects may arise and create negative externalities. Therefore, a proliferation of networks may create a situation akin to the 'tragedy of the commons.' The presence of additional participants will cause traffic jams and decrease the value of network membership to the other participants who are members of the network[9].

With respect to the creation of congestion effects, there is no doubt that the growth in data traffic may potentially overwhelm existing fixed line networks. This concern, however, has to be considered in the light of the facts presented by William Sharkey in his chapter "Representation of Technology and Production." Sharkey suggests that advances in optical networking technologies will allow optical signals to bypass the constraints at switching nodes, where optical signals have to be converted to electrical signals. Optical devices may lessen the impact of this constraint, which is responsible for the congestion effect. Hence, in future, if a number of small new entrants join the network and seek interconnection, the proliferation of firms, and networks, may not necessarily create a 'tragedy of the commons.'

A further issue examined by Liebowitz and Margolis is whether path dependencies play any important role in fostering inefficiencies, as is alleged in the now famous case of the QWERTY and Dvorak keyboards. Liebowitz and Margolis, however, have shown that there is not historical support for the common claim that the keyboard is an example of path dependent inefficiency. For telecommunications networks, if diseconomies on the supply side outweigh the scale economies on the demand side, then the unbundling of incumbent local monopoly telephone networks is likely to be economically justified. If, in the face of potential unbundling, local telephone monopolies are able to protect their entrenched positions, then customers on the incumbent monopolist's network might get locked into an inferior service, even though over time newer networks may provide better quality services and enhance consumer efficiencies. Efforts of either the incumbent monopolists or the regulatory authorities may frustrate efficient turnover. Here again, however, the competition from alternative networks – wireless networks, for example – limit the inefficiency that can be imposed by the 'locking in' of a self-isolating network. Thus, there are clear extrapolations from the existing literature on network effects to the contemporary telecommunications situation.

Brock's chapter usefully highlights the role that the Computer II has played, in the form of dampening prices for ISPs to connect to the PSTN, and in fostering Internet diffusion in the United States. Thus, the Internet has evolved as a separate network that operates in conjunction with the PSTN. If, however, this

[8] Even if interconnection is achieved at cost-based prices, the scale of entry matters for a network operator to benefit from the consumption externalities that are generated by entry of that firm into the network.

[9] Therefore, the issue is should several small networks be allowed to interconnect or a few large networks that can benefit from scale effects be allowed to do so.

high level of Internet diffusion is accompanied by congestion on the PSTN as a consequence of the Internet traffic flows to the final customer, then the benefits of Computer II are outweighed by the costs of congestion. There are, however, indirect network externalities that can be generated. The existence of both the PSTN and the diffusion of the Internet create a derived demand for new services. The customer connected to both the Internet and the PSTN can consume these services. The need to consider such indirect network effects then is also important.

2.3. Customer demand analysis

The chapter by Lester Taylor on "Customer Demand Analysis" provides a change in focus, away from the aggregate industry-level perspective of the first two chapters to the level of the customer in the telecommunications sector. At the consumer level, Taylor takes the presence of a network externality as axiomatic. Factors generating the demand for calling are exogenous. There are, however, endogenous effects that lead to the creation of dynamic network externalities, since once the exchange of information has taken place it creates the need for further exchange of information. A network externality can be construed as a demand-side economy of scale. Taylor brings out the need to think of the dynamic increasing returns process that propels these scale economies.

The distribution between static and dynamic network externalities is vital, but subject to considerable measurement problems, which have led to the emergence of only a small empirical literature on the subject, as compared to the empirical literature on the measurement of supply-side scale economies which is the topic of the next two chapters. As the types of available networks increase, with the addition of wireless and Internet connections to the customers' choice sets, the dynamic externalities increase. Therefore, the measurement problems affecting the concept of network externalities increase as well. Taylor's detailed analysis of the large literature on consumers' demand for access and usage, in conjunction with a consideration of some of the developments that have taken place in estimation methodologies, can help sort out the inherent difficulties in measurement.

Taylor's review of the demand modelling literature highlights the role that advances in discrete choice modelling have played in shedding empirical light on key questions. Allied with those are also the advances in simultaneous equation estimation techniques. Hence, as data on demand for access and usage of different types of network services become available, simultaneous estimates of fixed-line, wireless and Internet access and usage can capture the interdependencies between the demand characteristics for different networks. This type of modelling remains an agenda item that ought to shed light on consumption dynamics, as networks proliferate and converge. There are clear strategic implications of such modelling exercises for firms, as they begin to understand what motivates demand for different types of networks.

Taylor also highlights the related question of option demand. The benefit of telecommunications connectivity arose not only from the completed communications but also from the ability to do so. This issue of option demand gets more relevant as diffusion of the Internet and the mobile networks continue apace. Connectivity to the fixed-line network not only entitles the customer to make and receive voice calls, but also to the option of access to the Internet for the content contained therein. Simultaneous connectivity to mobile networks allows consumers enhanced option sets for accessing voice functionalities. Once wireless access to the Internet becomes widespread, the options available to the consumer will increase with the availability of alternative voice and content choices. As Internet based e-commerce becomes mobile networks based m-commerce, the options to both business and residential customers increase substantially. Then, basic telecommunications connectivity is expanded to include not just communicability, but also commercial capability. The choice sets for customers are increased still further.

How does the option demand interact with the dynamics of information exchange to alter consumer behaviour patterns? This important question requires research, at the theoretical and the empirical level, though the latter is more difficult since the data requirements are large. If consumer behaviour patterns do change substantially, then there are major implications for firms as well as policy makers. Take the issue of pricing. The way that prices are regulated for access to the Internet, and the way that prices are regulated for calls from fixed networks to terminate on mobile networks and vice-versa, vary across countries. These variations in prices are key drivers of consumers 'consumption behaviour given option demand and the dynamics of information exchange. If prices are set sufficiently low so as to encourage access, then customers' option sets increase substantially since there are more opportunities to make connections[10]. The availability of an increased number of potential connections can impact on the increasing returns process that underlies the dynamics of information exchange and increase the volume of digital signals flowing through networks. This enhanced volume can potentially cause congestion on the fixed-line network. If the supply-side economies of scale are robust, then the increased traffic volumes may be accommodated within one network[11]. Melvyn Fuss and Leonard Waverman, and William Sharkey deal with the consideration of supply side issues in their chapters.

[10] On this issue, we have remarked that the almost non-existent access charges for ISPs to connect to the local telephone networks in the U.S. have led to a substantial diffusion of the Internet in the U.S. Consumers in the U.S., thus, have considerably more choices than consumers elsewhere.

[11] An issue that the *Handbook* does not cover is what motivates firms in the telecommunications sector to make investments in the new technologies. A rich tradition in the literature, going back to Schmookler (1966), highlights the role of demand factors in influencing technology adoption. In the evolving telecommunications sector, how factors such as option demand and the dynamics of information exchange influence customer demand and thereby the potential deployment of new technologies by carriers remains an exciting topic awaiting research and investigation.

2.4. Econometric cost functions

The issue of whether the telecommunications industry is a natural monopoly or not has a long history. The fundamental institutional outcome, if a natural monopoly situation exists, has been the perceived need for regulation. In that context, the need for understanding what is an optimal market structure has generated a substantial literature, based primarily on Canadian and United States data, on scale, scope and efficiency issues. Melvyn Fuss and Leonard Waverman (Fuss and Waverman), judicially survey this literature, and the accompanying concepts, in their chapter "Econometric Cost Functions."

The first issue of concern is the distinction between three concepts: monopolisation, actual monopoly and natural monopoly. This distinction is highlighted in the chapter on competition policy by Daniel Spulber but relevant for this one too. The first concept is a process of behaviour by firms; the second concept is an outcome of that behavioural process in terms of market conditions; the third concept deals with the structural and technological conditions determining the cost-minimising number of firms in a particular market or industry. The last concept is what the current chapter deals with.

While in the decades from the 1960s to the early 1990s issues of natural monopoly and cross-subsidies were relevant, Fuss and Waverman identify access pricing and access conditions as the concerns of today. This observation implicitly goes to the heart of what is one of the crucial questions in the telecommunications field today: how many competitors should be allowed to enter the industry and on what terms and conditions? To understand the issue of how many requires detailed analysis of production conditions, which scale, scope and efficiency studies are meant to achieve. Thus, it is useful not to think of natural monopoly questions per se, but to ask: what is the structure of production like?

Having reviewed the available literature, and the techniques used in the studies, Fuss and Waverman reach the sobering conclusion that the existing literature does not really tell the policy maker what the permissible structure ought to be for a given market situation. There is a clear reason for this conclusion. Historically, the telecommunications sector has consisted of a monopoly operator that has operated both long-distance as well as local networks. The question of natural monopoly, at a system level, has arisen because it is assumed that an operator has both economies of scale in providing calls in general and that there are economies of scope, or multi-product economies, for that operator to produce two or more types of calls. Additionally, for producing any one type of call the costs of that producer are less than the total costs of two or more producers generating the same output. That is, at the system level costs are sub-additive. The empirical literature has never been able to test whether a stand-alone toll producer's cost function is different from that of a producer of both toll calls and local calls. Undertaking this test remains a challenge for the profession.

The Fuss and Waverman analysis raises the issue: what is the appropriate definition of the telecommunications industry for the purpose of such economic studies? Does it comprise all of the producers, collectively connected via their individual networks, in one large meta-network? Is it the long-distance operators who operate the trunk networks that may connect various parts of a country? Or is it the local exchange networks that operate in specific jurisdictions? Depending on what definition of industry is chosen, the econometric studies are carried out at that particular level. However, results obtained for one level of analysis, say the long-distance sector or segment, need not be consistent with the results obtained at another level of analysis, say that of the local exchange sector.

The appropriate definition of the industry is critical for several reasons. For the regulator, it is important to know what segment to regulate and what segment to keep free from regulation. For the prospective operator, industry definition is important because the efficient boundaries of a firm can be defined[12]. From an academic perspective, the definition of an industry is critical because it helps define the level of aggregation at which studies are carried out. The extrapolations to policy made from these studies are a function of the level of aggregation associated with each industry definition. If a study is carried out for one sector and the results are extrapolated to another, this can lead to wrong policy conclusions.

Fuss and Waverman's rather negative conclusions of the existing literature arise because academics have principally used aggregate-level data to answer questions that are inherently micro-level in nature. Take the controversies based on the U.S. Bell system data and captured in the works of Evans and Heckman (1984) and Charnes, Cooper and Sueyoshi (1988). Based on the use of a Bell system data-set, Evans and Heckman came to the conclusion that AT&T's nation-wide telephone system, consisting of the Long Lines long distance operations plus the local exchange operations, did not display subadditivity[13]. Therefore, a case could be made for breaking up the Bell system. Based on the use of exactly the same data set, Charnes, Cooper and Sueyoshi came to the opposite conclusion[14]. A problem in interpreting such diametrically opposing results is the level of aggregation at which the studies have been carried out. It is a flawed approach to club together data for long-distance and local operations into one data set and then to conclude whether the local or the long-distance sector is competitive or not. Such aggregated analysis can lead to potentially wrong policy decisions. Therefore, it is very important to use the right type of data for analysis.

[12] A related issue is the question of market definition, concerned with the set of competitors and their outputs.

[13] Shin and Ying (1992) use local operator level data for operators only in the state of California to reach similar conclusions.

[14] Röller (1990) comes to broadly the same conclusion as Charnes, Cooper and Sueyoshi.

Allied to the appropriate choice of industry segment to be analysed is the question of technique. In the core, the primary approach has been to use regression-based parametric efficiency estimation techniques. Such techniques, especially the translog variant, are flexible but can create problems because of underlying functional form specification oddities. Nevertheless, such techniques are extensively used in the recent literature. Another approach is to use operations research based techniques that do not, a priori, specify any functional forms. Not only is it feasible to estimate overall scale economies, but the precise optimal size of each unit of observation can also be estimated. Thus, data on the number of cost-minimising firms that can co-exist within a given industry can be ascertained. The use of this approach, involving the computation of firm or operator level scale economy parameters, augments the estimation of system-wide scale economy statistics.

The above property is useful for policy makers and for firm-level strategic decision-making. If the broad total size of an industry segment is known, given historical data on consumption and connections, these techniques permit derivation of optimal market structure since the size of the most efficient firm or operator can be estimated. Thus the number of firms that can co-exist, without any firms becoming inefficient, can be gauged. If the most efficient size is relatively small, there can be many operators within a given industry segment. For the prospective operator, such data reveal the level of investment necessary to be a viable player within a given industry segment. If the most efficient size of a firm is relatively small, this calls for proportionately less investment than otherwise would be the case. Given the potential policy and managerial prescriptions that can follow from using these new techniques, we hope that a significant body of work, exploiting new sources of data and new techniques, becomes available to shed continuing light on the dynamic yet vexatious problem of optimal market structure.

The question of what the supply-side economies are has to be considered in conjunction with the chapter on "Network Effects" by Liebowitz and Margolis. Network externalities are a demand-side scale economy characteristic. Such scale economies are a function of network size, but network size is also endogenous. Supply-side scale economies will only be generated if there is a large volume of calls being made on the system. The generation of demand-side and supply-side scale economies is a simultaneous process. From an academic research perspective, there are enormous benefits to combining the empirical literature of network effects with the empirical literature on scale and scope economies so as to understand the dynamics of the underlying increasing returns process.

Nevertheless, policy decisions have to be made in stages and sequentially. The supply-side scale economies can be such that they initially peter out at very small output levels. Then, given the generally burgeoning size of the total market, several entrants may be permitted to enter the market, as the market can accommodate many players. In fact, it is this assumption that led to the opening up the local exchange segment to competition in the U.S. 1996 Telecommunications

Act. Interconnection should permit a multitude of small entrants to eventually enjoy the system-wide demand-side scale economies that are feasible within the telecommunications sector. If some small entrants thereafter become more efficient because of firm-level managerial competencies, then the process of natural selection will mean that a consolidation of the sector might occur and large leaders emerge. The non-parametric approaches permit an ex-ante assessment of potential scale, or size, at entry. How, then, the dynamics of the increasing returns process play out in generating further economies is an empirical question.

2.5. Representation of technology and production

William Sharkey (Sharkey) in his chapter titled "Representation of Technology and Production" takes a deep micro-micro look at the underlying technology of telecommunications production. It is the last chapter of this section for several key conceptual and technical reasons. First, the micro-level issues that Sharkey deals with are now in the forefront of policy-making. Understanding what the supply-side economies are is useful in designing market structures. But the setting of prices requires an understanding of what it actually costs to produce a unit of output. In the erstwhile multi-product system based literature there have been considerable problems with the top-down allocation of joint and common costs, as well as estimating the cost functions. Therefore, the bottom-up cost literature has evolved so as to provide detailed estimates of what it might cost to provide the various services on an incremental basis[15]. This generation of ex-ante incremental cost structure estimates is very useful from a regulatory point of view. These estimates provide the basis for creating regulatory yardsticks against which subsequent performance may be measured.

Second, the bottom-up cost literature has evolved substantially in the U.S. because of the belief that the local exchange network, once believed to be the last bastion of natural monopoly, is prospectively competitive[16]. As Sharkey notes, the local competition provisions of the 1996 Telecommunications Act provide fundamental rights to prospective entrants. These rights are to interconnect with the incumbent operator at rates based on cost, to obtain unbundled network elements, and to obtain retail services at wholesale discounts. From an incumbent's point of view the knowledge of what costs actually are so that interconnect prices are set has been the raison d'etre for this approach. As countries around the world open up their monopoly networks to inter-

[15] The bottom-up approach is based on engineering models and primarily uses optimization techniques. We have remarked earlier on the use of non-parametric operations research based techniques for efficiency and optimal size measurement. These techniques provide a bridge to link the top-down and bottom-up approaches.

[16] This assumption is remarkably consistent with the position adopted a century ago!

connection by new entrants, the bottom-up engineering cost methodology will diffuse further.

Third, the practice of generating ex-ante cost estimates permits a forward-looking view to be adopted as to what a communications network ought to be like. While the adoption of a forward-looking view is enshrined in the 1996 Telecommunications Act as an economic principle, the implications of a forward-looking approach are more far-reaching. In taking a forward-looking view the engineering viewpoint converges with the economic viewpoint to create efficiencies.

Because of the technological change in progress, various alternate network structures are possible. Sharkey goes into details of the various network topologies. Each possible network topology can be created using alternate technologies. Thinking in terms of optimizing the traffic flows and using the appropriate technologies to then build the network prima-facie creates the least-cost traffic flow solutions. Networks are constructed to take into account local traffic and demographic idiosyncrasies, and this forward-looking approach helps operators plan for the best feasible solution.

Fourth, given the continuous technological changes taking place within the industry, the notion of a telecommunications firm is evolving. The evolving notion of what is a telecommunications firm leads to the re-drawing of market boundaries, a fact that Sharkey clearly notes. If the firm and market boundaries change, then the optimal size of a firm is also subject to change since the technological basis of production is in a state of flux. The change in the optimal size of a firm has implications for the overall structure of the industry. If the optimal size of the firm is small, then it is feasible to allow several firms to interconnect within a given geography.

Permitting a number of firms to enter and interconnect is an option to be considered in the light of Liebowitz and Margolis' comments on network externalities. They comment on optimal network size from a consumption point of view. If a prospective operator finds that an inability to enter at a large scale of operations leads to a denial of the consumption externalities then entry might not happen. Nevertheless, even smaller scale entry can be economically beneficial if optimal scale can be attained in the process. The volume of calls that terminate on the operator's network, however small, can generate access revenues that more than compensate for the loss of direct call revenues to be earned from having a large number of customers making calls from a larger size network. Thus, combining the ideas of Liebowitz and Margolis and Sharkey is important, from a competition policy point of view, in designing market structures and in deciding appropriate scale at entry.

There are further issues that emerge from Sharkey's chapter. As Sharkey notes, new technologies radically change the economics of transmission versus switching, thus expanding or contracting the scope of operations for a new entrant or an existing player. For example, a migration of scale economies to the switching function has occurred because of the use of common control features.

The above facts can have implications for the dynamic evolution of firms' capabilities within the telecommunications sector. In particular, the U.S. 1996 Telecommunications Act gives facilities-based entrants the right to interconnect with the incumbent operator on rates based on cost, for a lesser entrant to obtain unbundled network elements, and for an entrant which is a reseller to obtain retail services at wholesale discounts. It is thus feasible for a firm to enter the market as a reseller of calls that are purchased wholesale. The next stage can be to operate as a switchless service provider, using leased switches to operate the business. In the final stage, the entrant can acquire switches and provide switching capacity that can be the precursor to complete facilities-based competition where switching and transmission facilities are both provided[17]. Small-scale participation as a reseller provides the initial entry into the sector, and as efficiencies are achieved then growth to an eventual fully facilities-based player may be feasible. For policy makers, an understanding of bottom-up costs, and how such costs change, provides an indication of what types of firms might enter and in what number. Thus, bottom-up cost modeling provides a foundation for the analysis of market structures in a technologically dynamic context.

Finally, Sharkey provides examples of how bottom-up cost modelling has been applied in Argentina, Peru and Portugal. While it has been developed in the West, the methodology is robust enough to travel across borders and be applied in other contexts. It is not necessary to have advanced telecommunications network designs analysed with the help of this model. Even the establishment of basic networks, using simple and standard technologies, can be analysed using the approach. This way, the limited funds available to operators in developing countries can be wisely spent in designing the lowest cost network. Therefore, the Sharkey chapter is useful for policy makers from developing countries as well as developed countries.

[17] Sharkey, in a personal communication, further suggests that resale might not appear to be a viable long-run strategy for entry at least in U.S. local markets at the present time, though in the U.K. several service providers have entered the market for local and national long-distance calling. These service providers have taken away a large chunk of the incumbent monopoly firm's market share in the U.K. AT&T has complained about resale perhaps as an excuse for not entering U.S. local markets more rapidly. Nevertheless, AT&T's relative lack of success in entering local markets might be a function of perhaps aggressive entry deterrence strategies adopted by the incumbent local exchange carriers (ILECs) in the U.S. Resale nevertheless remains a valuable, and perhaps essential, transition strategy for new entrants. Switching assets are relatively scalable. Small switches are now available and adding line card modules can easily increase their capacity. Furthermore, digital loop carrier (DLC) technologies allow a single switch to serve a large geographic area at less than 100 percent penetration rates. Hence, competition in the switching element can be expected to occur in the reasonably near term future. The main problem with facilities based entry are the potentially large sunk costs in loop investment for wireline networks. Interoffice transport is by far the most competitive element, since entry by competitive access providers (CAPs) in urban markets is very well established.

3. Regulation

The second section of the *Handbook* contains a set of four chapters dealing with regulatory issues. Telecommunication has been and remains a highly regulated industry. However, the form of regulation has changed significantly over the past 30 years. The industry has progressively opened up to competition in all parts of the world, as new technology such as mobile telephony has been introduced and as the incumbent operator in many countries outside North America has switched from a public firm to mixed or wholly private ownership.

It is useful to identify three stages in the recent interactive processes involving regulation and the development of competition. For this purpose, we omit the precursors to the first stage of private or public monopoly. In the U.S., this period ended with the historic 'Kingsbury Commitment' in 1913, under which a series of *de facto* local monopoly franchises emerged under the leadership of AT&T. In European countries, such as the United Kingdom, the precursor to a publicly owned telecommunications monopoly was the development of a series of privately or municipally owned local companies. These were unified in the early part of the twentieth century into a monolithic organization at the national level, typically a department of government.

The regulatory arrangements during this first stage, which characterized the industry for the bulk of the twentieth century, can be dichotomized into a 'North American' model, based upon the independent regulation of private monopoly, and a 'European' model based upon direct regulation by governments of a public monopoly.

In the United States, the major instrument of regulation until the passage of the Telecommunications Act of 1996 was the 1934 Communications Act. This established the Federal Communications Commission (FCC). The FCC, in combination with state-level regulatory bodies and under the occasional supervision of the Courts, regulated an industry dominated by the vertically integrated Bell system. Until the 1950s, competition was virtually non-existent. The structure of retail prices, subject to break-even constraints, could thus be chosen with a high degree of freedom, without running the risk of competitive distortion. This structure took the form of relatively low monthly rental and local call charges, usually combined in a flat rate, and relatively high long-distance charges. Based on a combination of call types, the telephone companies were able to earn returns at least equal to their cost of capital. Low line rental charges encouraged high penetration rates, and hence the achievement of universal service objectives. A more explicit régime of cross subsidization, based on a detailed cost separations methodology, was developed. This was directed at mitigating cost differences between companies providing services in high cost and low cost parts of the United States.

The absence of competition, combined with public ownership of the industry, allowed governments operating the European model even greater discretion in

running the industry. The same government department was typically responsible both for the production and the regulation of telecommunication services. Public ownership mitigated the break-even constraint that characterized the American system. The European model typically embodied a structure of tariffs similar to that observed in North America, in the sense that line rentals are low (relative to cost) and long-distance call charges high. Within the relevant government department, the universal service objective of uniform and affordable tariffs could be achieved by direct government control of retail prices.

Brock, Kaserman and Mayo, and Woroch describe the development of the regulatory framework, within which competition in the United States was progressively introduced, in the chapters. It is useful, however briefly, to characterize by way of contrast the operation of the same processes in Europe and elsewhere.

Member states of the European Union (EU) operate in a federal framework. Under this framework the European institutions (the Parliament, the Council of Ministers and the Commission) enact legislation. This is then transposed into national laws and implemented by national regulatory agencies. After a number of member states opened up their markets to the competition in the 1980s, the European Commission promoted a package of legislation in the 1990s to establish a framework within which competition could develop. Crucially, a political commitment to liberalize telecommunications markets in the EU from 1998 was secured. Detailed implementation would rest with the national regulatory agencies. The regulatory framework did not, however, supersede European competition law. The Commission in its role of Competition Authority added considerable impetus to the process.

This set of regulatory arrangements can be broken down into provisions relating to licensing, interconnection and universal service. In addition, national regulatory authorities (NRAs) have typically imposed restrictions on the retail prices charged by the dominant operator, and also controlled the structure of prices. Examples are the limits on increases in line rentals and restrictions on the discounts which the incumbent might offer to the major customers. These customers were frequently the targets of entrants' marketing plans. However, the Commission has encouraged re-balancing of price structures away from cross subsidization.

This framework of regulation has been designed to meet the needs of a period in which the transition to competition was beginning. It has developed in a piecemeal fashion, and the instruments employed were often blunt. In some countries, notably the United Kingdom, the promotion of entry in the form of facilities-based competition was an explicit objective of policy.

Regulation in Europe is now about to enter a new phase corresponding to a period in which the market power of the incumbent is diminished. Regulation can, therefore, be less intrusive. The new régime also acknowledges the widening of markets associated with convergence. It is described as a framework for regulating not telecommunications but electronic communications. While main-

taining regulatory provisions aimed at the achievement of social objectives, such as universal service, the new framework is intended to bring regulation into closer convergence with competition law, focussing in particular upon firms in a position of dominance.

The regulatory framework elsewhere in the world has generally exhibited a similar progression, from regulation of a monopolist via a phase of early transition to competition, followed, where appropriate, by a relaxation of constraints on the historic operator. In Australia, this process has been accompanied by a reallocation of responsibility for the regulation of telecommunications to the competition authority (the Australian Competition and Consumer Commission). This body uses newly enacted provisions in competition law to control access to bottleneck facilities in network industries.

New Zealand represents the sole exception to the otherwise standard procedure of regulating the transition to competition to a regulatory agency operating under legislation that is additional to standard competition law. When the New Zealand government introduced full competition into telecommunications in 1989, it eschewed specific regulation and chose to rely instead upon its general competition law, the Commerce Act 1986. Extensive litigation ensued under the Act with respect to the interconnection arrangements between the incumbent monopolist, Telecom New Zealand, and a competitor, Clear. This finally culminated in a judgement delivered by the Privy Council in London in 1995[18]. This judgement did not, however, still the controversy. In 2001 the Government in New Zealand was giving consideration to the possibility of introducing telecommunications-specific regulation.

The previous paragraphs have provided a schematic view of the process by which regulation has developed to match the changing competitive circumstances. It is now appropriate to discuss in more detail the three key instruments reviewed in the regulatory chapters of the *Handbook*.

3.1. Price regulation

David Sappington (Sappington) in the chapter entitled "Price Regulation" discusses price regulation and incentives, with applications to telecommunications. The prime exemplar of such schemes are price caps, which are now widely employed by regulators throughout the world to control the retail prices charged by incumbent operators. A more recent, and interesting, development has been the extension of price caps to control the prices at which the incumbent sells network services both externally to competitors and internally through transfer prices to its affiliates.

Price caps are a member of a broader class of instruments of incentive regulation – mechanisms which provide operators with the opportunity to retain as

[18] See the chapter by Noam in this *Handbook*.

profit additional revenues or cost savings which occur, within a specified period, as a result of their own efforts. The alternative to incentive regulation is a régime in which the operator is entitled to recover in revenue whatever costs it incurs, including a fair rate of return on capital employed. For this reason traditional U.S. price regulation is often described as rate of return regulation, although it was in fact rarely a pure cost-plus system.

Sappington describes a variety of forms of incentive regulation, including banded rate of return regulation and various forms of earnings or revenue sharing. He also considers yardstick competition. This is a method under which a firm is rewarded on the basis of a comparison of its performance with that of a similar firm. However, price caps remain the most frequently adopted form of incentive regulation, both in the United States and elsewhere.

With a price cap, the regulator has to determine in advance the change (in telecommunications, normally a decline) in prices in real terms to which the firm will be subject. This typically involves an estimate made through benchmarking or the use of engineering cost models of the current level of efficient costs, and the forward projection of productivity gains or declines in unit costs. More sophisticated financial modeling may be employed with the aim of ensuring that, at the end of or throughout the price control period, the firm earns a rate of return on its projected capital base equal to the cost of capital. This formulation suggests a similarity between a price cap and rate of return regulation. However, there is the crucial conceptual difference that under rate of return regulation actual costs are reimbursed *ex post*, while with a price cap constructed in this fashion, the firm is entitled only to recover the regulator's *ex ante* projection of its costs. If it beats or falls short of that target its profits will be correspondingly affected. Sappington analyses the detailed ways in which price caps can be formulated and régimes in which operators are offered a choice between a price cap and rate of return regulation.

Considerable effort has been put into the empirical analysis of the effects on cost reduction and other variables of alternative methods of price regulation, normally using state level data in the U.S. Sappington concludes that incentive regulation appears to increase the deployment of modernizing investments, to increase growth in total factor productivity and to cut certain service prices. But there is little evidence of significant reduction in operating costs, and the effects are generally not dramatic.

3.2. Theory of access pricing and interconnection

Mark Armstrong (Armstrong) and Eli Noam (Noam) devote their chapters to, respectively, the "Theory of Access Pricing and Interconnection" and "Interconnection Practices." The importance of interconnection for the development of competition is reflected in the considerable progress in our understanding of the economic issues related to it, notably as a result of contributions by William

Baumol and Robert Willig[19] and Laffont, Rey and Tirole (1998). Armstrong distinguishes between two forms of the purchase of access or interconnection. In 'one-way' access, entrants need to purchase vital inputs from the incumbent but not *vice versa*. This form characterizes the stage of transition to competition. In 'two-way' access, all firms in the market need to purchase access to their rivals' subscribers to terminate calls. This régime has long characterized international telecommunications. This sector has traditionally required the co-operation of two established, and traditionally monopolistic, operators in the countries where the call is terminated and originated. It applies in a modified way to a more general multi-operator framework where competition is established.

In the model of 'one way' access, the key analytical and regulatory concern is that the incumbent will foreclose entry by setting access prices too high. In telecommunications, this familiar problem is often complicated by the presence of regulatory constraints on retail prices, in the form of unbalanced tariffs that depart from costs. These will have to be reflected in interconnection charges as an application of the theory of the second best.

A particularly influential formulation of access pricing with fixed retail tariffs is embodied in the so-called efficient component pricing rule (ECPR). According to this rule, efficient entry into the market is ensured by an access price set equal to the sum of the cost of providing access and the profit the incumbent loses in retail markets as a result of providing access to its competitors, also known as the 'opportunity cost.' Armstrong discusses the circumstances in which the ECPR is efficient. He also discusses the ways in which it requires amendments if there is differentiation between the incumbent's and the entrants' products, or opportunities for entrants to respond to high access charges by building their own facilities and by-passing those of the incumbent.

A more general framework within which to determine access prices is one in which the incumbent's retail and access prices are chosen simultaneously. This approach evokes the Ramsey principle. According to this principle, fixed costs are recovered via a mark-up inversely proportional to the price elasticity of demand of the final service. Access charges, differentiated from the final product into which they are an input, are inferred back from this principle. Armstrong further extends the discussion to consider circumstances in which the incumbent is subject to either no regulation, or partial regulation of its retail prices. Unsurprisingly, the optimal policy depends upon the assumptions made about the degree of competition in the market.

Summing up on 'one-way' access prices, he concludes that access prices should not be given too many tasks. In particular, they should not be asked to correct for distortions in retail tariffs. One of the advantages of eliminating distortions in the incumbent's retail tariff is that this permits cost-based access charges which do not require demand side information for their computation. These then yield

[19] Set out in Baumol and Sidak (1994).

correct signals to entrants contemplating by-pass. Ramsey pricing provides, by definition, an efficient configuration of retail and access prices, involving differential mark-ups. One version of the ECPR is valid when retail prices are constrained and other regulatory interventions excluded, but the rule has little relevance when the incumbent is free to choose its retail tariff.

If the fear in relation to 'one-way' access is foreclosure, the equivalent concern with 'two-way' access is collusively enhanced interconnection prices. These interconnection prices raise retail prices above efficient levels. In a simple model of 'two-way' access prices, exemplified by pre-competition international telecommunications, each firm has a fixed subscriber base. The issue then becomes one of non-co-operative or co-operative determination of termination charges. The outcome depends upon the circumstances in which the bargaining takes place; for example, on whether side payments can be made and whether termination charges must be reciprocal. In many circumstances, the outcome is inefficient and illustrates the process of double marginalization at work.

In the more general model, there is competition among networks for subscribers. This raises the question of whether such competition generates termination charges, which are close to the socially optimal, or whether the competitors are able to increase profits by negotiating particular access charges. Armstrong reviews a number of models bearing upon this situation. These differ in the assumptions they make about a firm's ability to employ non-linear retail tariffs, and to price discriminate between intra-network and inter-network calls. In some cases, a profit neutrality result is obtained: the access charge has no effect on equilibrium profit. Armstrong conjectures, however, that this is a relatively special case.

In summary, Armstrong shows that the 'one-way' interconnection literature has clarified the mapping between policy objectives and pricing principles, while the more recent 'two-way' issue still requires further development.

3.3. Interconnection practices

Eli Noam, in his chapter on "Interconnection Practices," focuses upon the different roles which interconnection has played in the development of competition, particularly in the United States, and upon attempts to apply pricing rules described by Armstrong. In the United States, interconnection began at the level of equipment through FCC decisions which allowed competitive apparatus owned by customers to be connected to the network. At the level of services, the long-distance sector was the first to become competitive, largely through the intervention of MCI in the late 1960s. Before the break up of the Bell system in 1984, the FCC had to approve a plan determining charges which long-distance operators would pay to local operators for call origination and terminations. After the

break-up, interconnection charges became the instrument of a (declining) cross-subsidy from long-distance to local operators. The 1996 Telecommunications Act codified new interconnection rules requiring the sale of unbundled network elements. The move towards flat rate access charges and the reduction in per minute access charges accelerated. The significance of interconnection in determining the viability of local competitors was increasingly recognized.

Similar developments were taking place throughout the world. Under the auspices of the World Trade Organization, many countries have committed themselves to ensure that their incumbent operators provide interconnection to market entrants at technically feasible points in the network on non-discriminatory and cost oriented terms. The scope of interconnection extended to cable networks, mobile networks and, increasingly, the Internet where backbone networks provided universal connectivity to subscribers anywhere in the world.

The pricing rules used in practice for interconnection cover a large range. They include 'bill and keep,' a reciprocal arrangement under which no payment is made to the operator to whom the call is transferred; various forms of average or fully distributed cost pricing, often employing historical cost asset valuation, and thus failing to provide the forward looking cost-based prices which would properly guide firms in their by-pass decisions; the application of price caps to control price levels for a range of interconnection services; and the Ramsey pricing and application of the efficient component pricing described above. Noam notes that historically interconnection has played different roles in the development of the industry. In its first stage it was used to maintain monopoly control. Later interconnection was used to open up the incumbent's network to new entrants. Supportive interconnection policies can accelerate competition, but their use should be time-limited. Noam questions whether the regulation of interconnection will always be necessary. If not, then relatively simple transitional rules can be employed. If it survives, then he foresees a continuation of micro management of the telecommunications industry and the perpetuation of a group of firms enjoying a client-like relationship with the regulator.

3.4. Universal service

The chapter by Michael Riordan (Riordan) is devoted to "Universal Residential Telephone Service." This is an evocative phrase denoting the general availability of affordable telecommunications services, the precise interpretation of which varies considerably over time and space. Justifications promoting universal service are both economic, based on the existence of network externalities, and political or social.

Riordan's analysis focuses upon telephone penetration in the modern United States. He examines the factors influencing local penetration rates, finding that

poverty is a major predictor of low penetration. On the hypothesis that prices are partly cost-based, he finds that higher average costs and longer loop lengths have negative effects on penetration.

The traditional means of promoting universal access is via a subsidy for the line rental charge, which is often bundled with local calls, financed by above cost prices for long-distance calls. Turning to normative issues, Riordan examines the efficiency losses associated with price distortion. He notes that in the light of these distortions, the FCC is moving in the direction of higher subscriber line charges combined with targeted subsidies. The wisdom of this policy should take account of the impact of economies of scale causing the marginal cost of connections to be low, but this does not appear to be an important justification for large price distortions to achieve universal service.

The significance of network externalities depends upon whether call externalities are important. If line externalities are mostly infra-marginal, it is likely that economic efficiency is promoted by targeting subsidies to marginal consumers who are most likely to generate call externalities. Riordan uses the literature on price discrimination to analyze ways of achieving this end by offering discounts to selected consumer groups and optional calling plans. He shows how the analysis of third degree price discrimination can lead to results similar to federal policies designed to subsidize subscribers in high cost areas in the United States.

The U.S. also targets universal service subsidies at low-income households. Several studies have been undertaken of the effect of these policies, but Riordan concludes that they provide little convincing evidence of impact. A study has also been undertaken of the effects of the high cost policy. It suggests that the subsidy is only weakly translated into lower prices, and hence probably ineffectual.

Riordan notes the relative paucity of empirical studies of the U.S. that would provide a firm foundation for universal service policies. This is even truer of other parts of the world, where a similar shift is occurring from universal to targeted subsidies, associated with the re-balancing of the tariff and the elimination of cross subsidies as competition develops. In Europe, mechanisms have been created to ensure that the burden of meeting universal service obligations is shared in a fair way among competitors. These have involved forming estimates of the net cost of universal service. This is normally defined as the losses which the universal service operator incurs, calculated on an avoidable cost basis, in supplying customers whom it would not serve on a normal commercial basis. Estimates of the cost of universal service, defined in this way, have been small. They are typically equal to one or two percent of total fixed line telecommunications revenue. This figure would, however, clearly increase if, as has happened in the U.S., the obligation were expanded either to include the delivery to all households of more advanced services or to provide subsidized services to key public facilities such as schools and libraries.

3.5. Interaction among regulatory institutions

The approach adopted by the authors of the chapters described above has been normative. An objective function has been implicitly or explicitly defined, and alternative policies evaluated in relation to the degree to which they achieve it. The institutional framework through which regulation has been effected is not examined. There are, however, two additional and related issues. The first involves analysis of the interplay among economic interests in shaping regulation. The second involves a more detailed analysis over the interaction among regulatory institutions.

Impetus towards the first method of analysis, known as the 'positive theory of regulation' came from early work by Posner (1974) and Peltzman (1976). In their approach, regulatory intervention is a service supplied by regulatory institutions and demanded by those affected by regulation. In equilibrium, a particular form of regulatory intervention emerges, which is provided at an equilibrium price. One of the key insights of this approach is that consumers of regulated services are unlikely to be able effectively to realize their demand for intervention, because of the free-riding problem. Regulated firms, on the other hand, have a more concentrated interest in purchasing regulatory services, which they are able to make effective. As a result, capture of the regulatory agency by the firm is likely to occur. An interesting case is that of firms as users. Traditionally, firms in the U.S. have paid higher local service charges than households have paid. However, over the last decades telecommunications-intensive businesses have benefited more than households have from the advantages of competition.

The second issue concerns how regulatory institutions in telecommunications can be designed to ensure an appropriate balance between the interests of consumers and the interests of producers. It is assumed that competition in the market in question is not effective. Thus some form of regulatory intervention is required. Absent such intervention, consumers would be the victims of excessive pricing. However, when such regulation is present, investors in sunk productive assets run the risk of seeing prices depressed to a level which covers the variable costs of production but fails to cover total costs. If the firm's assets are sunk, it will continue to produce in these circumstances. What is required, therefore, is a balanced set of arrangements. These should protect consumers and provide comfort to investors. Levy and Spiller (1996) and Spiller and Vogelsang (1997) have analyzed the circumstances in which both these objectives can be achieved. Potential contributors to a solution, in the form of an effective system of checks and balances, are legislative bodies, the bureaucracy, and the judicial system.

This analysis highlights the potential importance of the governmental and regulatory structure. It is common for a sector to be regulated by more than one authority. In some cases, several regulators can regulate a single firm. In other cases, competing firms can be regulated by different regulators or under different rules. This practice leads to cooperation and conflict among regulatory agencies.

In the U.S. context, all the cases described above occur. For example, a local exchange company is regulated by a state public utilities commission for its end-user services, and for interconnection with other network and service providers serving the same state. However, it is regulated by the FCC for interconnection and access services used to produce interstate communications. With respect to competition issues, such as mergers, a regulated telecommunications company can be subject to scrutiny by the state regulator, the FCC and the antitrust authorities.

The division of labor between the state regulators and the FCC is governed by the 1934 Communications Act and the Telecommunications Act of 1996 and by general constitutional principles. The main such principle is that of federal pre-emption. This principle says that the FCC is in charge of issues that involve more than one state. The FCC can preclude state regulators from regulating an issue if interstate matters are involved. However, many issues involve both interstate and intrastate matters. In such cases, preemption is warranted if the interstate matters cannot be separated either physically or through accounting procedures. The principle of federal preemption can be contrasted with the subsidiarity principle which is applied in Europe. The subsidiarity principle decentralizes regulation by giving authority to the lower level of a federal hierarchy whenever that can be done.

The division of labor between different regulatory levels can be justified in terms of differential abilities to regulate. Federal regulators can make use of economies of scale from dealing with many cases. They can resolve conflicts arising from matters crossing local boundaries. At the same time, local regulators can have more specific information and tailor decisions to local concerns. The 'laboratory of the states' can also find new solutions for new problems more quickly and experiment with them before they are adopted on a grand scale. This has happened with interconnection regulation, where collocation and unbundling were developed by progressive U.S. states, such as New York and Illinois. Equally, overlapping regulation creates opportunities for firms to engage in forum shopping by choosing to appeal to the regulator most likely to accede to their wishes[20].

[20] Overlapping regulation slows down the regulatory process. New laws in the U.S. always work slowly. It is an integral part of the U.S. legal and regulatory system to create delay. This happened following passage of the 1996 Telecommunications Act. The apparent winners were the U.S. incumbent local exchange carriers (ILECs), for whom delay means the prolongation of high profits. Delay served the interests both of the ILECs and of those competitive local exchange companies (CLECs) that have problems getting ready for the big competition. Cable TV companies needed time to upgrade their networks and to ready themselves for the higher quality service levels required by telephony. They were also calmed by a reduced threat of ILECs invading their TV territory. Long-distance companies also have less desire to enter local markets in a big way now that long-distance access charges have been reduced. They found that local network entry is painfully costly and time-consuming. At the same time the ILECs experienced the flip side of delay because they did not get permission to enter long-distance markets before opening up their local networks to full competition.

4. Competition

The third section of the *Handbook* is about competition. The past two decades have been interesting times for the telecommunications sector. One of the most interesting features has been the spread of competition. This has started with terminal equipment and long-distance services, and moved on to mobile and local services. As shown in the chapters by Brock and Woroch, the path to competition has not been a one-way street. In particular, local competition had already existed a century ago, after Bell's original patents expired. As we have noted earlier in this introduction, it was the failure of this competition, due to the lack of interconnection, which led to entrenched monopoly regulation and legal barriers to entry. The chapters by Noam and Armstrong show how interconnection regulation has been the key to network competition throughout the world, while universal service provisions analyzed by Riordan have been instrumental in countries like the U.S. to pacify incumbent carriers' opposition to competition. The current section starts out with an analysis of competition policy in general and then focuses on competition in specific areas. This analysis focuses more on policy concerns than on structural issues, several of which have been covered earlier in this introduction.

4.1. Competition policy

In the U.S., the telecommunications sector has, for almost a hundred years, been subject to both competition policy and regulation. During this time, the jurisdictional separation between these two policies has not always been clear. Daniel Spulber's chapter on "Competition Policy in Telecommunications" is, like Hausman's on the mobile sector, guided by a reasoned aversion to regulation. Spulber argues that, for a substantial amount of time in the recent past, U.S. telecommunications policy reversed the role of regulation and competition policy. The regulator, FCC, implemented competition policy, while the antitrust authorities and Judge Greene, who was in charge of the AT&T divestiture, did a lot of the regulation.

In Spulber's view, industry specialists such as the FCC are not impartial enough to pursue competition policy. Meanwhile generalists such as antitrust authorities and courts are not expert enough, and not subject to enough pressure by interested parties, to regulate. The interchange of roles between competition policy and regulation has also occurred in countries like the U.K., where the regulator has a strong role in competition policy, and in Australia, where the competition authority has become the regulator. A major difference to the U.S. case may lie in the long case history that has guided U.S. regulatory and antitrust authorities in certain directions. Meanwhile regulatory and antitrust agencies in other countries are much more open. Nevertheless, finding the right balance between regulation and competition policy is an important task of

public policy design, as competition in telecommunications progresses on a worldwide scale.

Spulber brings out how competition policy is constrained by regulation of telecommunications. The competitive problems are monopolization, leveraging and mergers. Spulber finds that the enormous changes of the telecommunications industry in the last two decades have eliminated most hurdles to the feasibility of competition. The growth and diversification of markets and the convergence of the media have reduced natural monopoly issues. Twenty years ago natural monopoly was doubted only for long-distance markets but clearly assumed to hold for local telephony[21]. Today, it has become unclear to what extent natural monopoly prevails at all. Unsustainability is only perceived as a problem in connection with cross subsidies imposed by regulators, while contestability is viewed as a chance if entrants are helped by access and interconnection regulation. The power of potential competition has been replaced by the power of actual competition.

The main question, when moving from a regulated or state-owned monopoly to competition, is if the former monopolist has sufficient advantages over entrants to be able to stifle competition. The answer depends largely on the presence and significance of sunk costs. Spulber conjectures these to be small. While this hypothesis might be controversial, there does not appear to exist good empirical work to determine the extent of sunk costs in the telecommunications industry. Spulber has powerful arguments in favor of the small size of sunk costs. The most convincing of these is the fast growth and technical change of the industry. While true, this seems to stop at the local loop. Most analysts concentrate on the local loop as the basis for both natural monopoly and sunk costs in telecommunications. The local loop has been remarkably stable and long-lived. It is also the most costly item in the network. Thus, Woroch, in his chapter on local network competition, gives examples that might suggest the presence of economies of scale and sunk costs in parts of the local network.

One consequence is certainly that more empirical research is needed. The econometric estimates of cost relationships that are presented by Fuss and Waverman are plagued by statistical and data problems. They also show that the evidence for natural monopoly in the local network is weak. As we have earlier commented, the relevant results were obtained based on system-wide estimation. At the same time, the proxy cost models described by Sharkey come to the conclusion that local network costs are a declining function of network density. These models are based on local loop level data. Nevertheless, these data are ultimately the products of forecasts and simulations of production for assumed topologies.

If the econometric models were to be relied on, there would be a strong case for pursuing infrastructure competition throughout the network. If the engineering

[21] See the chapter by Fuss and Waverman in this *Handbook*.

models were to be relied on, infrastructure competition in local networks would make sense only in very dense networks, where economies of density are exhausted. In other areas, competition should be restricted to some form of resale of services or of network elements. There is, however, a different policy consequence from the empirical ambiguity that might be drawn at the current stage of our knowledge. This policy approach starts with the observation that, conceptually, there could exist two important ranges of natural monopoly.

In the first range, the natural monopoly property can be weak. In this range, economies of scale and scope are almost exhausted and sunk costs tend to be small. In this situation, competition is likely to be beneficial, because it leads to pressure on costs, prices and innovation. Competition is also likely to occur here, because in most telecommunications markets demand is moderately to strongly inelastic. Thus, we can expect duplicate network investments, associated with some cost inefficiency and excess capacity, but possibly lower prices than under regulated monopoly. This is the case of long-distance services, mobile telephony and local business services in downtown areas in industrialized countries. In the second range, the natural monopoly property can be strong. In this range, there are economies of scale and sunk costs are large. In this situation, competition would most likely be wasteful and not self-sustaining. However, if we allow and do not subsidize competition it is unlikely to come. This is the case of local networks in rural areas.

There is little harm in allowing competition in telecommunications in general. This consequence is, of course, well known at this time. It is practiced in wide parts of the world. There is, however, another less-appreciated consequence. It is that our yardstick for measuring the competitiveness of telecommunications markets should be the amount of actual competition that occurs if markets are fully opened. Rather than asking the question as to whether a market is a natural monopoly, we ask what amount of competition occurs? If there is no competition in spite of open markets, the strong version of natural monopoly may hold. Whether or not that is the case, such a market requires further regulation. If competition occurs and spreads out quickly, natural monopoly may be absent or only weakly present. In this case, deregulation is warranted. If some competition occurs, but the incumbent former monopolist remains dominant, we have the intermediate case. Here softer regulation may be in order, but not yet full deregulation. Competition thus becomes the discovery process for the nature of production conditions and for guidance on the policy of deregulation. The entrepreneurial experiments that are put in place will generate the evidence as to what the evolving market structure is going to look like. A process of natural selection will weed out the inefficient firms that enter and retain the efficient ones. If efficiency is accompanied by large size, then a potential conclusion is that economies of scale exist.

Two caveats are in order. First, an incumbent's anti-competitive behavior could interfere with the competitive outcome. Spulber argues that predatory pricing,

leveraging, foreclosure, tying and cross-subsidization are unlikely to occur, unless the regulatory process provides incentives or opportunities for such behavior. It is thus the second caveat, regulatory distortions that seems to be more important. Rate-of-return regulation supposedly leads to cross-subsidization of competitive services by the regulated services of a utility, while various entry barriers have been responsible for tying and foreclosure. The other side of cross-subsidization is cream skimming. This may have been responsible for the first episode of facilities-based competition by competitive access providers in the local network. Cross-subsidization, at the same time, creates regulatory foreclosure in the subsidized areas.

In order for actual competition to fulfill the function as a discovery process for the natural evolution of market structure, regulation has to be made competitively neutral. This requires three tools: the abolishment of legal entry barriers; a functioning incentive scheme for end-user regulation; and, a competitively neutral scheme of access regulation. The interaction of these three tools and the ensuing difficulties in their application can be followed through the three competition cases analyzed in this volume. They are long-distance competition, the mobile telephone market, and local network competition.

4.2. Long-distance competition

The chapter on "Long-distance Competition" by David Kaserman and John Mayo (Kaserman and Mayo) shows how legal entry barriers have suppressed otherwise feasible competition in the U.S. long-distance market for a long time. It also shows that access regulation to the local network was necessary to make long-distance competition work. A functioning incentive scheme for end-user regulation was not part of the required tool set for a long time. On the contrary, inefficiently regulated long-distance charges helped long-distance competition to emerge in the first place. Incentive regulation was only established almost 20 years after long-distance competition had started. Then, however, it provided for a smooth transition from regulation to competition. In 1995, the general public hardly noticed when AT&T's deregulation took place.

Competition in the U.S. long-distance market has provided the yardstick for competition in other parts of the telecommunications sector. It has demonstrated that competition works and can be maintained over objections that it interferes with "network integrity" and the natural monopoly property of the network. It was competition that demonstrated the weakness of the natural monopoly property in long-distance services just in time before growth of fiber optics could have led to a revisionism and could have been used as an argument in favor of stronger natural monopoly. Although Faulhaber (1987) had demonstrated the existence of substantial excess capacity in the long-distance networks, few have used this as an argument against competition.

U.S. long-distance competition also paved the way for the regulation of access to the local network. At first, rivals access was of lower quality than what AT&T provided to itself. In a tedious process, equal access was established as part of AT&T's divestiture, and the subsidies for local services contained in long-distance access charges were gradually reduced. In fact, not only did access to the local network provide the key to the feasibility of long-distance competition. It was also the decline of access charges that explains most of the decline in long-distance service prices.

Kaserman and Mayo show how long-distance services have evolved into a normal, albeit highly concentrated, unregulated market. They provide substantial evidence that this market functions reasonably well. This evidence has to be seen in light of the controversy during much of the 1990s over the competitiveness of long-distance telephony. This was spurred by the observation that, AT&T's declining market share notwithstanding, long-distance profit margins seemed to be increasing. Kaserman and Mayo lead us through an array of arguments and evidence to contradict this view. Since this discussion was shaped in adversary regulatory and judicial processes, and since experts of long-distance competition have taken sides, the reader is asked to read this assessment with an open mind. The authors' arguments are provided in a very systematic manner that guides readers through the various methodologies of assessing the amount of competition. They help readers judge competitiveness and the reasoning behind deregulation of long-distance telephony.

In spite of the evidence of functioning competition in U.S. long distance telephony, entry barriers have not been fully eliminated. The former regional Bell operating companies (RBOCs) are allowed to enter interLATA long-distance markets only if their local markets are deemed sufficiently open to competition. This demonstrates a fear on the part of the U.S. legislator that the regulatory tool of access regulation cannot be fully depended upon. Fears of foreclosure, and exploitation by incumbents of the advantages stemming from the ability to offer the full array of telecommunications services, are behind this policy. If these fears were justified it would be a particular problem for countries outside the U.S., where the former PTTs or other monopoly providers have remained vertically integrated. In these countries the carrot of entry in long-distance markets cannot be used to induce such firms to improve conditions of access and interconnection in local markets. Those countries therefore may have to resort to stringent enforcement of local access and interconnection rules and to the development of some local competition before they deregulate long-distance services.

Kaserman and Mayo include international telephony from the perspective of competition in the market for U.S. international calls, but avoid the morass of international telephone markets, including issues of accounting rates, call-back services and the like. The tendency of accounting rates to lead to potentially inefficient outcomes is treated on a theoretical level in Armstrong's chapter on access charges. International calling is a strongly growing area that, however,

becomes more like domestic long-distance calling, as time progresses. Its importance is likely to be dwarfed by Internet developments, which will be assessed in a subsequent volume of the *Handbook*.

4.3. Mobile telephone

Mobile telephony has been one of the immense drivers of the telecommunications industry. Its growth has been phenomenal, so much so that mobile subscribership will soon surpass that of fixed telephone lines. Jerry Hausman (Hausman), in his chapter on ''Mobile Telephone,'' impressively works out how the growing importance of the sector affects the benefits that flow from good regulatory and business decisions and the costs from bad ones. He shows how the FCC's decision to postpone the introduction of cellular telephony by about a decade has cost the U.S. economy (and, in our view, the rest of the world as followers) billions of dollars. He further demonstrates that taxing mobile telephony has had very substantial distortionary effects, much larger than raising ordinary taxes. He also shows how, in the U.S., state-level price regulation was accompanied by substantially higher prices than the prices that would have prevailed under unregulated duopoly competition as in other states. All these observations are derived with sound and clearly laid out methodologies. They lead Hausman to conclude that, except for spectrum allocation, the government should stay out of mobile telephone markets.

Hausman's chapter includes experiences from countries in Europe but his analysis is largely restricted to the U.S. How does it travel abroad? What makes most countries (all but most of the OECD countries) in the world different from the U.S. with respect to mobile telephony is that they start out with a much less developed fixed network. As a consequence, particularly in poorer countries, mobile telephony can provide, for the first time, full telephone coverage of the country at affordable costs and within a short period of time. The relationship between mobile and fixed services in poor countries is thus substantially different from the situation in the U.S. Mobile telephony is even more important in those countries. Hausman's hypotheses on the undesirability of regulation could, therefore, hold even more strongly there. In addition, complementarity and substitutability between mobile and fixed networks may differ between poor and rich countries. In poor countries, there will be a distinct regional complementarity arising from the fact that, without mobile networks, large sparsely populated areas would not be served at all. In densely populated areas, there will be more substitution. This will hold, in particular, if the fixed network providers cannot keep up build-out of their networks with growing demands. Otherwise, throughout the world, one main competitive issue is the extent to which wireless services will be a substitute for wireline services. This is treated by Woroch, because this is how local network competition could well gear up in the future.

4.4. Economics of spectrum auctions

Mobile telephony's most distinctive input is radio spectrum. In the past, such spectrum was awarded in cumbersome administrative processes, called 'beauty contests,' or through lotteries. Lately, however, auctions have taken over as a revenue raising process of allocation. The chapter by Peter Cramton (Cramton) on "Spectrum Auctions" links the issues of competition *in* the market and competition *for* the market. Spectrum auctions are, in the first instance, competition for the market. They, however, determine competition in the market, because the spectrum set aside for a particular purpose limits the total capacity available for that particular set of services. Then the identity, number and distribution (geographically and in terms of bandwidth) of winning bidders determines the extent of competition in the market. Experience of competition with more than two participants is starting to indicate that third and fourth competitors have significant effects on competitiveness of mobile markets. An early paper by Selten (1973) on imperfect competition, evocatively sub-titled "where four are few and six are many," suggests that competitiveness may further grow, as the number of mobile competitors grows to five or six. These developments of mobile markets provide for natural experiments of the effects of entry into markets and should, therefore, attract doctoral students in empirical industrial organization.

What Cramton's analysis brings out very clearly is the learning from bad experience in auction design. Spectrum auctions have, from the beginning, been influenced and accompanied by economic analysis. Still, economists did not always anticipate bad outcomes. They did not warn the New Zealand government that second price auctions could lead to very low auction revenues if the number of bidders becomes very small. They did not warn the FCC that small company privileges would be competed away in auctions with small companies only (that could be backed by large companies). This failure actually shows that the auction market worked! Cramton therefore highlights the problems that auction design incurs and how they can be solved or have been solved by the FCC. For example, he shows that collusive bidding could be achieved if bidding increments are allowed in odd $ amounts. This way, bidders could use the last few digits of their bids as codes to provide signals to their competitors. Allowing simple multiples of pre-specified bid increments only could undermine this ability.

Another important issue highlighted by Cramton is the relationship between the number of licenses offered for 3G spectrum and the number of incumbents. These incumbents need the 3G spectrum to stay in business and are therefore willing and able to out-compete potential entrants. Thus, in order to attract additional bidders and increase auction revenues, the number of new licenses has to exceed the number of existing licenses by at least one or two. Otherwise, the incumbents will collude among themselves and bidding will result in dismal revenues. At the same time, offering too many licenses could bring down auction

revenues because competition in the market would be viewed by auction participants as too intense to be profitable. The auctioning authorities thus have to strike the right balance between competition in the markets for wireless services and auction revenues for the government.

4.5. Local network competition

Glenn Woroch's concluding chapter on "Local Network Competition" synthesizes the issues treated in all the other chapters in this *Handbook* because the local network is so much the core of the telecommunications sector and so intertwined with its other segments. At the same time, it is the local network where competition is the weakest and regulation most persistent. Woroch shows that part of the resistance to competition in the local sector can be traced to bad experience with competition a hundred years ago, when Bell's patent expired and duplicate local networks emerged in many U.S. cities. The problem that destroyed competition at the time was the lack of interconnection. With interconnection, enforced by regulation and with resale rules and network unbundling in place, competition in the local network is no longer doomed. It is, however, developing from a small base and currently stands at a level that, in competition policy terms, would likely be viewed as monopoly. Woroch, however, shows that the current growth of competition is so impressive that the assessment may change within a few years. During this time, the scope of the local market is likely to change as well. Geographically, 'local' and 'long-distance' concepts will become blurred. What is likely to survive and gain in significance is the distinction between the access network and transport network[22].

The main bottleneck is now, and will remain for some time, final customer access. It may have strong natural monopoly properties. In contrast, the transport network has strong natural monopoly properties only in sparsely populated areas (or in poor countries more generally). The scope of local markets is also likely to change by including other services that are currently provided by firms in related but not yet integrated markets, cable TV and mobile telephony, in particular. Cable TV is already integrated with local telephony in the U.K., but that model does not travel well to countries with existing cable TV systems, because it is based on the integration of telephony and TV at the time of the network buildout. However, due to the integration of high-speed Internet connections with telephony, in other countries such as the U.S. cable TV has become a close substitute to the high-frequency portions of copper telephone loops used for broadband services.

Woroch devotes substantial space to the present and future competitive relationships between wireline and wireless telephony. If the substitutability between mobile and stationary telephones increases in the future it will have

[22] The latter includes switching.

enormous implications for the competitiveness of local markets. Because of its huge customer base, the competitiveness within the mobile sector would spread to the wireline sector. In particular, the significance of the local loop would be undermined, because the mobile network does not have a local loop that is sunk[23]. This would particularly affect rural areas, where fixed local loops are so costly to provide. Competing mobile carriers, who need to have nationwide coverage in order to offer full mobility, would in all likelihood serve such areas. Thus, in order to continue competing, local wireline carriers would have to take a write-off on their terrestrial loops in remote areas. Universal service policies could interfere with this danger for some time. This development could also prevent geographic price differentiation from happening. The other side of this development could be that the competitive position of wireline networks is strengthened in agglomeration areas, where spectrum for wireless services is scarcer and where the demand for high bandwidth is particularly great.

We believed that by now (in 2001) competition would ride in high gear. That optimism was premature. What took twenty years in the simple long-distance market may take a long time in complicated local markets. Competition in local telephone markets has a long way to go, in the U.S. and elsewhere. So far, competition in telecommunications has been uneven. It is most pronounced in the long-distance and mobile sectors. But it has barely scratched the surface of local telephony. However, it is coming from all sides, directed at the incumbent local exchange carriers (ILECs). ILECs may well continue to dominate their own turf in the long run the way AT&T has continued to dominate long-distance markets over the last two decades. They hold impressive advantages over rivals in the form of ubiquitous two-way networks and existing customer relationships with almost 100% of the population.

ILECs, however, will ultimately face formidable competition and be squeezed from at least two sides. First, wireless services are changing from being complements to being substitutes for landline services. Even if wireless local loops do not take hold immediately, they will come sooner or later. Thus, the wireless and landline services will eventually occupy the same market. Within current wireless markets, competition is already proceeding at a fast pace, and that is accelerating through provision of further spectrum in auctions. Second, broadband services will be in increasing demand, putting cable TV companies, competitive access providers and specialized internet backbone providers into intensive competition with DSL services of traditional ILECs.

In the end, the ILECs' reluctance to embrace new entrants by offering them quick and reasonably priced network access may turn out to be their biggest mistake. This reluctance has triggered tough new rules by the legislature. The ILECs might postpone their application but that cannot be prevented completely. It triggers facilities-based entry by network providers who will not go away again.

[23] Not even a wireless local loop (WLL) is a sunk loop in the same sense that a terrestrial loop is.

Similar dynamics have occurred in other countries, where the reluctance of incumbents to enter into negotiated interconnection agreements, while postponing competition, spurred entry helped by the regulators.

Once duplicate local telephone networks exist, their capacity needs to be used. The closer one comes to the final user, the more do networks assume natural monopoly properties. This means that duplication close to the users bears the danger of excess capacity and fierce competition[24]. This could, in the long run, trigger collusion between ILECs and CLECs. Before this happens, in the local area there are probably three episodes in what we have above viewed as the third stage of competition and regulation. At first, both retail and interconnection markets are regulated. The slower competition progresses in local markets the longer this episode will persist. In the second episode retail markets are successively freed from regulation but interconnection remains regulated. This episode will last as long as vital parts of ILEC networks represent bottlenecks that have to be used by competitors. In the third episode, antitrust replaces regulation at both market levels. A main obstacle to reaching this point in the U.S. is the open-ended universal service policy that was necessary to win over Congressional support for the competitive provisions of the 1996 Telecommunications Act. To some extent, New Zealand has reached the third stage. The U.K. has just postponed plans to enter the second stage. The U.S. is still in the second stage for long-distance and mobile competition, but quite firmly in the first stage for local telephony. Thus, for a change, the U.S. may be able to gain speed by learning from others.

As Spulber, and Kaserman and Mayo, Hausman and Woroch observe, the arguments and evidence on competition in telecommunications markets are indeed strong. The main monopolistic niche is the local access network. Even that is under attack from substitutes in the form of cable TV and mobile networks. The more the media converge the closer becomes their level of competition. For this to happen, however, regulation has to be restricted to some core areas, such as the auctioning of spectrum rights and the functioning of interconnection. As this happens, competition policy assumes an increasing role.

5. Conclusion

The chapters that follow in the *Handbook* describe a revolution in the telecommunications sector over the past 20 years or so. This revolution embraces fundamental changes in technology, market structure, and the way in which analysts, policy makers and regulators think about the industry. The earlier view of telecommunications as a natural monopoly has now given way to one in which

[24] On the other hand, as Taylor has suggested, in his chapter in this volume, additional demand may well be generated to meet the supply of capacity available, as the *dynamics of information exchange* described by Taylor in this *Handbook* trigger additional uses for the capacity in place.

almost all parts of it are seen as susceptible to some form of competition. Depending upon the activity in question, competition may take the form of access to the incumbent's network, leasing of unbundled loops or other assets, or even resale of services. Increasingly, however, it takes the form of duplication of facilities.

Technological change has played a major role in creating these new opportunities. In particular, developments in wireless technology, which are proceeding at an extremely rapid pace, have not only made possible the development of a range of new services, but have also created the possibility of generating competitive constraints on the monopoly of highly concentrated wireline networks. Equally, the development of cable telephony has created a competitive local loop that in many countries, particularly those with a high rate of cable penetration, already has access to a majority of homes. There remain questions about the feasibility of rival broadband networks to the home. At present, however, the momentum is in favor of significant reductions in the concentration of markets. Yet, the incumbent operators are taking significant strategic steps to retain their positions, and the process of market structure evolution seems exceedingly dynamic and interesting.

Time-honored and in many cases inefficient forms of regulation have acted as a constraint on the development of competition. These developments do not mean that the spread of competition will inevitably remove the need for scrutiny by regulatory and competition agencies of telecommunications markets. Bottlenecks will remain, particularly those associated with call termination. Small numbers of competitors, creating the opportunity for tacit collusion will characterize parts of the network. But many industries have special characteristics of this kind, without being treated as special cases.

Our advice to policy makers is to embrace these competitive developments. This applies both to developed countries and to developing countries where the absence of pervasive existing networks creates particular opportunities for the deployment of a variety of new technologies, some of which particularly lend themselves to use in a competitive environment. The chapters that follow show that competition is possible, but also that its introduction should be based on careful application of the appropriate economic concepts. They also show that competition is not inconsistent with the attainment of the broader social objectives that the telecommunications industry has typically pursued.

References

Baumol, W. and G. Sidak, 1994, The pricing of inputs sold to competitors, Yale Journal of Regulation, 11, 171–202.

Chandler, A., 1977, The visible hand: The managerial revolution in American business, Cambridge, MA: Harvard University Press.

Charnes, A., W. W. Cooper and T. Sueyoshi, 1988, A goal programming/constrained regression review of the Bell system breakup, Management Science, 34, 1–26.

Evans, D. S. and J. J. Heckman, 1984, A test for subadditivity of the cost function with an application to the Bell system, American Economic Review, 74, 615–623.
Faulhaber, G. R., 1987, Telecommunications in turmoil: Technology and public policy, Cambridge, MA: Ballinger Publishers.
Kuhn, T., 1962, The structure of scientific revolutions, Chicago: University of Chicago Press.
Laffont, J.-J., P. Rey and J. Tirole, 1998, Network competition: I. Overview and non-discriminatory pricing, RAND Journal of Economics, 29, 1–37.
Levy, B., and P. Spiller, 1996, Regulation, institutions and commitment, Cambridge: Cambridge University Press.
Peltzman, S., 1976, Towards a more general theory of regulation, Journal of Law and Economics, 19, 211–240.
Posner, R., 1974, Theories of economic regulation, Bell Journal of Economics, 5, 98–115.
Rohlfs, J., 1974, A theory of interdependent demand for a communications service, Bell Journal of Economics, 5, 16–37.
Röller, L.-H., 1990, Proper quadratic cost functions with an application to the Bell system, Review of Economics and Statistics, 72, 202–210.
Schmookler, J., 1966, Invention and economic growth, Cambridge, MA: Harvard University Press.
Selten, R., 1973, A simple model of imperfect competition: where four are few and six are many, International Journal of Game Theory, 2, 141–201.
Shepherd, W. G., 1983, Concepts of competition and efficient policy in the telecommunications sector, in E. Noam, ed., Telecommunications Regulation Today and Tomorrow, New York: Harcourt, Brace and Jonanovich, Inc.
Shin, R. T. and J. S. Ying, 1992, Unnatural monopolies in local telephone, RAND Journal of Economics, 23, 171–183.
Spiller, P. and I. Vogelsang, 1997, The institutional foundations of regulatory commitment in the UK: The case of telecommunications, Journal of Institutional and Theoretical Economics, 153, 607–629.

SECTION I – STRUCTURE

This Page Intentionally Left Blank

Chapter 2

HISTORICAL OVERVIEW

GERALD W. BROCK

The George Washington University

Contents

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

Before 1845, long distance communication in the United States was slow and expensive despite government efforts to develop an extensive postal system serving all parts of the country. While the postal system was widely used for business, the basic charge of $0.25 per sheet of paper for letters sent more than 400 miles put routine communication out of the financial reach of ordinary workers who earned wages of $1.00 per day. The Post Office Acts of 1845 and 1851 simplified the rate structure and drastically reduced the prices to a flat rate of $0.03 for letters sent anywhere in the country. The three-cent basic rate remained generally in effect for over one hundred years until it was raised to four cents in 1958 (U.S. Census, 1975, p. 19; John, 1995, pp. 159–161). The three-cent flat rate stimulated a massive increase in the volume of mail, from an average of three letters per person, per year in 1840 to an average of sixty-nine letters per person, per year in 1900 (John, 1995, p. 4; U.S. Census, 1975, p. 8, 805). The flat rate also accentuated the previous practice of subsidising the sparsely populated areas of the country through charges on the densely populated areas.

The Post Office Act of 1845 coincided with the initial private development of the electric telegraph. While the Post Office did not place any legal restrictions on the development of the telegraph, the telegraph companies were required to respond to Post Office prices and policies for financial viability. The telegraph was a "fast letter" that could not compete in price with the regular mail, but could charge a premium for time sensitive communication. By 1870, the number of Western Union telegram messages was two percent of the number of letters while Western Union's revenue was thirty-five percent of the Post Office revenue (computed from U.S. Census, 1975, p. 788, 805). The invention of the telephone in 1876 added a third product to the nation's rapidly expanding information infrastructure. By 1900, many people had a choice of three communication technologies: inexpensive universally available mail service, expensive fast telegrams for time-sensitive written communication, and widely available local voice telephone service together with limited high-cost long distance voice telephone service.

While much of the world combined the three technologies into a single government agency, the U.S. retained a government Post Office and privately owned telegraph and telephone companies. Nevertheless, the privately owned telephone company sought regulation as a defence of its monopoly position. By emphasising politically controlled monopoly rather than market-oriented competition, the U.S. private telephone company developed many characteristics similar to the European publicly owned telephone companies and to the U.S. Post Office. By the 1960's, the domestic telegraph service had declined to secondary importance, but the telephone monopoly had achieved a thoroughly protected position. State laws and regulatory policies protected it from competition. Prices were designed to produce enough revenue to cover total cost (including depreciation and return

on capital), but the prices of individual services were determined by political considerations rather than the cost of service. The extensive implicit cross subsidies built into the price system provided a political rationale for excluding competition in any particular service because limited competition could take away profits used to subsidise favoured users. The telephone company accepted end-to-end responsibility for the service and strictly limited the ability of any other company to connect to its network.

The stable monopoly equilibrium was upset by the spread of time-sharing computers during the late 1960's. In contrast to the earlier batch processing or process control computers in which all inputs and outputs were physically close to the computer, time-sharing computers allowed computer service to be provided to distant terminals. The natural way to utilise the existing infrastructure to facilitate distant communication was to use terminals connected to telephone lines to access the computers. However, that required connecting the telephone line in some way to the computer, a violation of American Telephone and Telegraph Company's strict non-interconnection policy. Although AT&T developed methods of supplying computer communication needs, computer users were dissatisfied with the limited set of standardised products provided by AT&T. Efforts to provide freedom for computer users to develop and utilise specialised communication products with the telephone system created a limited form of competition, and a consequent need to define the boundary of the monopoly. Defining the boundaries of the monopoly became extremely complex and contentious once the implicit boundary of complete end-to-end monopoly telephone service was overturned. Several attempted boundary lines were overturned by the efforts of competitors to expand their business opportunities. The AT&T divestiture eventually shrank the boundary of the monopoly from the entire telephone network to only local exchange service (excluding customer premises equipment and inside wiring) in 1984.

After the divestiture, a separate challenge to the remaining monopoly arose from the rapid growth of data communications. The development of inexpensive Ethernet systems for local data communications sparked the growth of local area networks (LANs), first to link computer workstations and then to link personal computers. The LANs were private systems for internal communication and not directly competitive with the public telephone network. However, when the LANs were then connected to the public Internet (outside of telephone company control), they became a substantial alternative communication system that provided indirect competition to the telephone companies. The success of the non-regulated decentralised network of networks that made up the Internet provided support for the idea that monopoly restrictions at the core of the telephone network could be removed without public harm. The Telecommunications Act of 1996 changed the policy structure to a system of interconnected networks with no monopoly core, but direct competition with the incumbent local exchange telephone companies remains very limited.

2. Development of regulated monopoly telecommunication

2.1. The early telegraph and telephone industry

Telegraph development in the United States began in 1844, under the sponsorship of the Post Office. After operating an experimental line, the Post Office declined to purchase the patents from inventor Samuel Morse and allowed the telegraph to be developed as a private enterprise. Morse then licensed several different entrepreneurs to build individual lines along the East Coast and west to St. Louis. A dispute over royalties with the license holder for the western line led to a breakup of the patent organisation and efforts by many parties to build telegraph lines outside of the control of the Morse patent. Court efforts to shut down the lines that were not licensed by Morse were unsuccessful and telegraph building became an industry with quite free entry. Capital costs were relatively low, and building techniques were primitive in the early days because of lack of experience and lack of fundamental understanding of electricity. However, the Morse Code and apparatus designed to utilise that code could be used even in very poor conditions and lines were built rapidly through the sparsely populated but large U.S. territory.

During the 1850s the complementary nature of railroads and telegraph was utilised by Western Union to establish a major position in the U.S. industry. Western Union developed contracts with the railroads in which the railroad would be a partial owner of the telegraph and would give Western Union exclusive use of rail right of way for telegraph messages. Western Union agreed to provide priority to railroad operational messages and to transmit them without charge. During that time frame the railroads were expanding rapidly and were finding a great need for communications in order to operate safely and efficiently. Many railroads were single-track lines that required knowledge of train movements at a distance in order to avoid accidents. Furthermore, the increasing complexity of railroads made it necessary to utilise extensive communications in order to manage the far-flung enterprise. In particular, the Erie Railroad, under General Superintendent Daniel McCallum, established a formal operational control system with hourly and daily telegraph reports in order to control the rapidly expanding system.

Initially, Western Union was one of many telegraph companies serving the United States. The company moved toward dominance of the industry during the U.S. Civil War, 1861–1865. Just before the outbreak of the war, Western Union secured a government contract to build a telegraph line to the West Coast. With the outbreak of the war that line became extraordinarily profitable and could command rates of up to $1.00 per word because of the great time savings in communicating over telegraph versus any other method of communicating over the very long distance. Western Union lines were almost entirely in the states that remained with the Union. Many other telegraph lines extended across the

boundary between the Union States and the Confederate States and were cut at the outbreak of hostilities. Furthermore, the general manager of Western Union was appointed as Director of the Military Telegraphs for the Union Armies. He directed the construction of extensive military telegraph lines during the war and those were absorbed into Western Union's system after the war. By 1866, the year after the war ended, Western Union was the dominant telegraph company in the United States and provided a nation-wide telegraph network (Brock, 1981, pp. 73–83). Western Union expanded rapidly in the decade following the Civil War, increasing its offices from 2,250 in 1866 to 7,072 in 1876, while increasing its miles of wire from 76,000 to 184,000 over that same time period, and increasing its revenue from $4.6 million to $9.1 million (U.S. Census, 1975, p. 788).

With rapidly increasing telegraph demand, an incentive was created to transmit multiple messages over the same wire simultaneously. In 1872, Western Union purchased a duplex patent providing the ability to transmit two messages simultaneously on one wire. In 1874, Thomas Edison developed a quadruplex apparatus. Other inventors continued the effort to improve the capacity of telegraph lines. Elisha Gray, an electrician for Western Electric, which was partially owned by Western Union, developed a method of carrying up to fifteen conversations by using different tones on the telegraph in 1876. That same year, Alexander Graham Bell filed a patent application titled "Improvement in Telegraphy" that contained a vague speech claim. Both Western Union and Bell considered themselves the inventor of the voice telephone. After unsuccessful initial negotiations, both began establishing telephone services in 1877, with the first telephone switching exchanges in 1878.

In 1878, the Bell system was reorganised with professional management and Theodore Vail was hired to develop the company. The next year, he reached an agreement with Western Union in which Western Union and Bell agreed to pool their patents in telephone under the control of Bell, while Bell would stay out of the telegraph business and send all telegraph messages it received by phone to Western Union. Western Union also agreed to finance twenty percent of the Bell research and was to receive twenty percent of Bell's rental on telephones during the 17-year agreement.

The 1879 agreement between the Bell Company and Western Union led to a strict patent monopoly on voice telephone service by Bell while Western Union continued to develop the telegraph business. The Bell patents were upheld as controlling all voice communication over wire, not merely the specific apparatus originally patented to transform voice into electrical signals. Had the courts ruled for a narrow interpretation of the patents, they would have become largely irrelevant because greatly improved telephone instruments quickly superseded the specific apparatus. With the broad interpretation, Bell gained a full monopoly until the expiration of the basic patent in 1893 (Brock, 1981, pp. 89–104). It enforced the monopoly vigorously and succeeded in shutting down any companies that attempted to provide telephone service outside of Bell control.

During the Bell patent monopoly, telephone service was largely confined to local areas, and, therefore, was complementary to the long distance service provided by Western Union telegraph, rather than directly competitive. In that time period Western Union continued to grow rapidly as telegraph communications became a critical component of the rapidly expanding U.S. economy of the late 19th century. Between 1876 and 1893, Western Union increased its annual messages from 19 million to 67 million, while it increased its net revenue from $9.1 million to $23 million in the same time period (U.S. Census, p. 788). Western Union messages increased at a compound annual rate of 7.5 percent per year during that time period, while the average revenue per message dropped from 49 cents per message to 35 cents per message.

The Bell System initially focused on building telephone exchanges in the centre of the cities, and on developing short toll lines. The technology of the time was not yet developed to build true long distance lines, but it was gradually improved to increase the range of voice conversations. By the expiration of the patent monopoly in 1893, the Bell system had installed 266 thousand telephones or 4 telephones per 1000 of population. The Bell System was approaching the size of Western Union by that time.

In the years following the expiration of the Bell patent monopoly, telegraph messages carried by Western Union remained the dominant form of long distance communication in the United States. However, Bell put great efforts into developing long distance voice technology and expanding its system. Western Union had a dominating price advantage for very long routes, but a lesser advantage for shorter routes. For example, by the turn of the century, the price of a basic long distance telephone call (3 minutes) was twice the rate for a basic telegram (10 words) between New York and Philadelphia, but the price of a basic telephone call was 10 times the price of a basic telegram between New York and Chicago. At that time, voice telephone service was unavailable across the country. When full cross-country telephone service first became available in 1915, it was priced at $20.70 for a basic New York to San Francisco call, while a telegram between those points cost $1.00 at the same time (U.S. Census, pp. 784, 790).

After the expiration of the basic Bell patents, many new companies entered the telephone business in the United States. Initially, entry occurred in places that had not been served by the Bell System. Bell had concentrated its efforts on the major cities and had left large portions of the country without telephone service. There was extensive demand for telephone service in small towns and rural areas. Commercial companies and co-operative arrangements were formed to serve these areas. In many cases, the equipment was primitive and of low quality. At that time, the Bell company had purchased the Western Electric Equipment Manufacturing Company and had signed an exclusive dealing arrangement in which Bell agreed to purchase only from Western Electric and Western Electric to sell only to Bell licensed companies. Consequently, the independents were

required to develop their own equipment or find an alternative source other than the Western Electric Company that had previously dominated telephone equipment manufacturing.

A very large number of independent companies were formed in different parts of the country and they expanded rapidly. As they began to develop, they also began supplying competition directly to the Bell companies in areas already served by Bell. With the beginnings of competition, prices were reduced and growth accelerated. The early competition was non-interconnected competition in which network effects were a dominating factor in the choice of a telephone. Individuals could not make calls from an independently owned telephone to a Bell licensed telephone. Consequently, the choice of company depended on which company had the most subscribers with whom one was likely to wish to communicate. In many cases, that simply meant the most total subscribers, but in some cases a smaller network that contained people for whom communication was desired could be more desirable than a large network containing a random mix of the population.

With the entry of new telephone companies, and a reduction in prices, the growth rate of telephone service accelerated from an average of 7 percent per year during the monopoly period to an average of 25 percent per year. The growth rate of Bell owned telephones rose even while the company was facing competition with non-Bell owned telephones. The period of intense competition between 1893 and 1907 caused the total number of telephones to increase from 266 thousand at the end of the monopoly period to over 6 million by 1907. At that time, the U.S. had an average of 7 telephones per 100 population (U.S. Census, 1975, pp. 783, 784).

The development of the telephone in the United States during the monopoly period was comparable to the early development in the government-owned companies of Europe during the same time period. In both cases, the companies focused on telephone development in the major cities, with relatively high prices and slow growth and little effort to explore the limits of demand in the smaller areas. However, the U.S. rapid development under competition was not matched by the government-owned telephone operations of Europe. During the competitive period, U.S. telephone development increased faster than telephone development in Europe, with particularly notable differences in the small towns and rural areas. For example, at the beginning of the 20th century, Berlin and New York City had almost equal telephone coverage with approximately 2.5 telephones per 100 persons in the population. However, at that date, the rural state of Iowa had twice the telephone development of New York City with 5 phones per 100 population (provided by a total of 170 different systems within that state), while rural districts in Germany had less than one tenth the development of Berlin with two-tenths of a phone per 100 persons in the population (Brock, 1981, p. 144; Holcombe, 1911, pp. 418–438).

2.2. Regulation and monopoly

During the monopoly period, the Bell system had financed its growth largely through retained profits. With the rapid growth of the competitive period, and the reduced profitability because of price competition, the company needed outside capital in order to finance its growth. Early in the 20th century, the Bell system began regular interaction with the New York financial markets. In one particularly important transaction the company placed a very large set of convertible bonds with J.P. Morgan and Associates. The bonds failed to sell and Morgan and Associates took an active role in reorganising the company. In 1907, Theodore Vail, who had left the company after initial leadership to pursue other business interests, was appointed as president with the support of J.P. Morgan. Vail began a vigorous strategy to restore the market power of AT&T, the parent company of the Bell System.

Vail developed a three-part strategy to restore market power: merger with the independent telephone companies; increased emphasis on fundamental research and the purchase of outside patents; and a welcome to regulation.

The merger strategy was one being used by many other companies in the U.S. at that time. The U.S. was undergoing its first great merger wave in which large numbers of companies were consolidated in order to eliminate competition and gain potential advantages of economies of scale and scope. J.P. Morgan and other New York financiers were deeply involved in the merger wave, and Vail's merger strategy was comparable to that being used in other industries. In 1909, AT&T purchased a controlling block of Western Union stock and Vail was chosen president of Western Union as well as AT&T, creating an effective merger between the telegraph and telephone companies. Vail also began purchasing individual independent telephone companies. In 1911, an effort was made to consolidate the industry through a complete purchase of all independents, but the negotiation broke down because of disagreements among the many independent telephone companies and Vail continued to purchase individual companies, reducing the market share of independent telephone companies from 49 percent to 42 percent between 1907 and 1912.

As Vail purchased independent telephone companies, the remaining ones began to feel more isolated and vulnerable. Bell had the only long distance network of the time, but some independents were planning a separate long distance network to connect with other independents across the country. However that plan was threatened as individual potential participants in the network moved into the Bell companies. The remaining independents went to the Department of Justice to seek action against the Bell system under the Sherman Act. Bell was planning to consolidate a number of independent telephone companies in Ohio into the Bell system. The Attorney General expressed an opinion that such a consolidation would be a violation of the Sherman Act and began discussions of a settlement.

In 1913, the Bell system reached a settlement (known as the Kingsbury Commitment) in which AT&T agreed to sever its ties with Western Union, allow interconnection with the remaining independent telephone companies, and refrain from acquiring any more directly competing companies. The Kingsbury Commitment established a durable telephone industry in which the Bell system controlled most of the telephones and provided service in the cities while a large number of independents continued to serve the rural areas. After the agreement, the independents that were directly competing with Bell companies exchanged territories so that each had a geographic monopoly. The Bell and independent companies connected to a single long distance network and to each other to create a system in which any telephone could be reached from any other telephone.

The second portion of the strategy was to greatly increase basic research in the Bell System. Bell had been performing research and had generated many patents. However, those patents were not nearly so dominating as the original telephone patent and competitors were able to find ways to operate outside of the Bell patents. The early 20th century was a time of rapid technological advance and two innovations in particular created possibilities for competition with the telephone system. Marconi's early radio telegraph service could not at that time carry voice, but it provided an alternative to wires for communication and could potentially compete with voice communication after further technical refinement. The vacuum tube, developed by Lee DeForest, was in an early stage of development but was protected by patents and showed promise for telephone amplifiers and other new developments.

The Bell system greatly expanded its research capability, multiplying the number of engineers in Western Electric's research and development unit by a factor of five between 1910 and 1915. AT&T also created the organisation, later known as Bell Laboratories, by changing its research focus from modest improvements in engineering methods to more fundamental research. A 1911 report stated:

> "It was decided to organise a branch of the engineering department which should include in its personnel the best talent available and in its equipment the best facilities possible for the highest grade research laboratory work ... A number of highly trained and experienced physicists have been employed, beside a group of assistants of lesser attainments and experience" (quoted in FCC, 1939, p. 187).

AT&T purchased the patent rights to the DeForest vacuum tube in 1913 and continued developing it for use in telephone repeaters and radio. Prior to the practical development of the vacuum tube there was no useful way to amplify voice circuits. Long distance communication was limited and expensive because special construction, with very low loss lines, was required in order to extend the distance over which voice could be carried. The vacuum tube was also critical to the further development of radio. AT&T's control of the vacuum tube patent

allowed it to negotiate complex patent cross licensing agreements with General Electric, RCA, and Westinghouse in order to gain access to technologies developed by those companies in return for use of the vacuum tube in radio broadcasting and other uses. The patent agreements provided a method to draw strict industry boundaries that prevented outsiders from entering specific sectors while dividing the telephone and radio business among the incumbent companies.

The third element of Vail's strategy was to embrace regulation and use regulation as a substitute for market forces. There was extensive concern about the power of monopolies in the early twentieth century and efforts to use either antitrust policy or regulation to limit their power. Vail recognised that regulation could be a way of preserving monopoly power in justifying a system without competition. He began advocating regulation in the 1907 annual report and continued seeking it in a number of public forums. In a later speech Vail stated,

> "I am not only a strong advocate for control and regulation but I think I am one of the first corporation managers to advocate it. It is as necessary for the protection of corporations from each other as for protection to, or from, the public" (Vail, 1915).

Vail's advocacy of state regulation to eliminate competition followed the pattern of an earlier effort to use city franchises to eliminate competition. In the territory of Southern Bell Telephone Company, the company made private commitments to business organisations to secure their support for city regulations that allowed the Bell company to purchase the competing independent and obtain an exclusive city franchise for providing telephone service (Weiman and Levin, 1994).

Federal regulation of telecommunication was established with the Communications Act of 1934. That Act was written in quite general language and was administered by the Federal Communications Commission (FCC), a body created by the Act. The FCC initially consisted of seven members (later reduced to five members) who are appointed by the President and confirmed by the Senate. The commissioners serve for fixed terms and cannot be removed by the President during their term. Consequently the Commission is more independent than executive branch agencies that are under the direct control of the President. The Commission creates its own executive structure and appoints its own staff. Many of the routine orders of the Commission are issued by senior staff members under authority delegated to them by the Commission while major policy decisions are made by vote of the entire Commission.

The Communications Act of 1934 centralised and strengthened previous regulatory authority over both common carrier and broadcasting functions. The FCC was assigned responsibility for allocating spectrum used by entities other than the federal government and for granting licenses to that spectrum. Initially, much of the activity of the Commission was concerned with assigning radio and later television broadcasting licenses and establishing rules for the use of

broadcasting stations. However, the allocation of spectrum and the assignment of licenses have played increasingly critical roles in the regulation of common carrier telecommunication.

The early FCC was an ideal regulatory agency from AT&T's perspective. It provided very little control or restriction on AT&T's interstate rates and activities but it did help prevent competition from arising. The FCC did not undertake a formal investigation of interstate telephone rates until the 1960's, some thirty years after the Commission was formed. In earlier periods it operated under a system called 'continuing surveillance' in which no formal investigation was undertaken or orders issued but staff members reviewed AT&T's operations in an informal way and made suggestions.

The most important protective action of the FCC in the early period occurred in the late 1940's. At that time microwave technology had just been developed as a result of wartime work on radar and communications systems. Because of the government sponsorship of the wartime work (most of it undertaken at MIT's Radiation Laboratory and AT&T's Bell Laboratories) much of the technology was in the public domain. In particular, MIT had sought to make the technology easily available by publishing a multi-volume series on the microwave innovations it had developed during the war (Buderi, 1997, p. 251). Microwave technology provided a means of carrying either many telephone conversations simultaneously or a broadcast video signal. At about the same time that microwave technology was becoming available, early television stations were being established, and they sought a means to communicate among themselves and transmit programs across a network. AT&T's existing lines were inadequate for that purpose.

Microwave communication did not require rights of way as previous wire line communication did, but microwave required licenses from the FCC because of the use of spectrum. Initially the FCC freely granted experimental licenses for microwave communication. Philco achieved the first operational domestic microwave link when it completed a Washington to Philadelphia line in the spring of 1945. Western Union followed with a New York to Philadelphia link and Raytheon planned a transcontinental system. General Electric and IBM planned a microwave system for business data communications. However, the FCC deferred the question of permanent allocations for microwave during the experimental period. In 1948 the Commission adopted AT&T's position that only common carriers ought to have access to the microwave frequencies. In that decision the Commission allowed television networks to continue to build temporary microwave communication systems because of AT&T's shortage of facilities, but it required the networks to abandon their private facilities when AT&T facilities became available. Furthermore, the Commission declined to require interconnection between AT&T and Western Union microwave facilities for transmitting video programming. The result of the restrictive policy was to effectively reserve the new field of television transmission to AT&T (Brock, 1981, pp. 180–187).

The late 1940's microwave controversy was significant in enhancing AT&T's monopoly power. The new technology provided a means of entry for companies other than AT&T while the television demand provided a source of new demand. Thus it was not necessary to remove customers from AT&T in order to utilise a microwave system. A critical part of breaking down monopoly power is the development of facilities in special niches that can later be expanded into other parts of the market. In this case, the FCC prohibited the development of specialised facilities for video transmission and allowed AT&T to utilise the regulatory process to enhance its market power.

3. The shrinking boundary of regulated monopoly

3.1. Competition in customer premises equipment

AT&T prohibited customer owned attachments to its network from the early days of telephones. Restrictions were included in customer contracts prior to regulation and were incorporated into tariff language in 1913. AT&T enforced its 'foreign attachment' ban vigorously though unsystematically. In at least one case the telephone company prohibited customers from putting a cover on the telephone directory because a cover was an attachment to the telephone book which it considered part of the telephone system. However, AT&T allowed many unauthorised recording devices to be attached to the network through haphazard procedures in enforcing its restrictions.

The first important legal precedent limiting AT&T's control of attachments to the network came from the Hush-A-Phone case. The Hush-A-Phone was a cup-like device that snapped on to the telephone instrument to provide speaking privacy and shield out surrounding noises. It was a passive non-electrical device that directed the speaker's voice into the instrument. The Hush-A-Phone was sold for about thirty years before the case came to the FCC. Throughout that time AT&T made occasional efforts to stop sales of the device. In 1948, the Hush-A-Phone inventor filed a complaint with the FCC alleging that AT&T was interfering with his business. The FCC ruled in favour of AT&T, utilising the principle that "telephone equipment should be supplied by and under the control of the carrier itself." Bernard Strassburg, the FCC staff council in the Hush-A-Phone case, later explained his reasoning in recommending against Hush-A-Phone as follows:

> "It was the conviction of the FCC and its staff that they shared with the telephone company a common responsibility for efficient and economic public telephone service and that this responsibility could only be discharged by the carrier's control of all facilities that made up the network supplying that service. Such control included not only transmission, switching, and the subscriber station used for basic end-to-end service. It also had to extend to any equipment a subscriber might attach to or interface with the basic service. Only by this comprehensive type

> of control could the quality, safety, and economies of network performance and design be assured. The Hush-A-Phone, by itself, posed no threat of any real consequence to the performance of the basic network. Nevertheless, authorisation of its use could set a precedent for other, less benign attachments which individually or cumulatively could degrade network performance to the detriment of the public (Henck and Strassburg, 1988, pp. 38, 39).''

Tuttle did not accept his loss in the FCC proceeding as final and appealed the decision to the appeals court, where he won a clear victory and set an important precedent on the limitations of the telephone companies and the FCC to restrict privately beneficial use of the telephone network. The court delineated the customers right to use the telephone in privately beneficial ways so long as they were not publicly detrimental and ruled that the telephone company could not interfere with that right. The 1956 court decision stated in part:

> ''The question, in the final analysis, is whether the Commission possesses enough control over the subscriber's use of his telephone to authorise the telephone company to prevent him from conversing in comparatively low and distorted tones . . . to say that a telephone subscriber may produce the result in question by cupping his hand and speaking into it, but may not do so by using a device that leaves his hand free to write or do whatever else he wishes, is neither just or reasonable. The interveners' tariffs, under the Commission's decision, are an unwarranted interference with the telephone subscriber's right reasonably to use his telephone in ways which are privately beneficial without being publicly detrimental'' (Appeals Court, 1956, p. 269).

While the 1956 Hush-A-Phone court decision concerned a minor device with no significance for the overall telephone network or competition in that network, it established a legal principle that later was interpreted as a ''harm to the network'' principle. That is, the telephone company could only prevent devices that caused harm to the network and could not arbitrarily prevent all attachments to the network. However, initially AT&T interpreted the decision narrowly and revised its tariffs to allow the Hush-A-Phone and very little else. The FCC did not object to AT&T's narrow construction at that time.

The significance of the 1956 court decision did not become clear until twelve years later when the Commission was ruling on another disputed device known as the Carterfone. The Carterfone connected mobile radio telephone systems to the telephone network by utilising a cradle into which an ordinary telephone handset could be placed. The Carterfone transmitted voice signals from the mobile radio transmitter into the telephone handset and converted the voice signals received from the handset into radio signals for broadcast to the mobile radio telephone. AT&T threatened to suspend telephone service to any customer who used the Carterfone and Carter sought assistance from the courts and the FCC. In 1968 the FCC ruled in favour of Carterfone and expanded its interpretation of the earlier Hush-A-Phone precedent by requiring AT&T to file new tariffs that allowed all devices that did not cause harm to the network.

The FCC's change of position in the late 1960's was caused by the development of time sharing computers. Time sharing computers required the joint use of

telephone lines and computers to allow remote users to communicate with a central computer. A strict application of the restrictive connection rules would have prohibited communication with computers over telephone lines or would have required AT&T to supply the computer. AT&T had begun supplying teletypewriter terminals and modems as the terminating equipment for data lines and then allowing the interconnection of an AT&T supplied modem with customer owned computers. Both AT&T and the FCC viewed the Carterfone case in the context of the computer-communications boundary issue. Both were looking for a way to open the network to connection with new types of equipment for computer usage. AT&T chose not to fight the Carterfone concept and saw advantages for AT&T in allowing a wider variety of terminals to be connected to the system. AT&T could not supply all of the specialised devices that might be required by computer users but it also believed that centralised control of the telephone network and of the signals used to control that network was necessary.

After the Carterfone case, AT&T allowed connection of terminal equipment through a special device, known as a protective coupling arrangement, which was designed to maintain control of network signalling. However, as new companies came into the market they sought direct connection to the network without paying the telephone company for a device they considered unnecessary. During the 1970's extensive controversies occurred over the appropriate degree of terminal competition. New AT&T leadership sought to limit terminal competition and restore the traditional end-to-end service responsibility of the telephone company. In 1973 AT&T Chairman John DeButts publicly challenged the FCC policies and sought the support of the state regulatory commissions to limit competition in terminal equipment. Several states supported AT&T's anti-competition stance, with policies regarding terminal equipment becoming a contentious issue in relationships between the FCC and AT&T, and in relationships between the federal government and state government authority to manage telecommunication policy. The long established jurisdictional line that interstate services were subject to the federal government supervision and intrastate subject to state governments' supervision did not provide clear guidance in this case because the customer's telephone and other terminal equipment were used for both intrastate and interstate calls. The legal controversies were settled with a victory for the federal government that allowed the FCC to promulgate rules that could not be contradicted by state governments (Appeals Court, 1976, p. 787).

The controversies of the 1970's allowed a messy regulatory situation in which individual customers purchased some unregulated terminal equipment and the telephone companies provided some as part of a regulated service. The structure was clarified and improved by the FCC's 1980 decision known as Computer II that concluded the Commission's second attempt to distinguish the dividing line between regulated communications and unregulated data processing. In the 1980 decision the Commission concluded:

> "We find that the continuation of tariff-type regulation of carrier provided CPE [customer premises equipment] neither recognises the role of carriers as competitive providers of CPE nor is it conducive to the competitive evolution of various terminal equipment markets. We find that CPE is a severable commodity from the provision of transmission services... the separation of CPE from common carrier offerings and its resulting deregulation will provide carriers the flexibility to compete in the market place on the same basis as any other equipment vendor" (FCC, 1980, pp. 446, 447).

The Computer II decision completed the CPE story by shrinking the boundary of regulatory policy to exclude CPE. With that decision the long-standing question of what attachments to the telephone network must be considered part of the regulated industry was definitively answered. The regulated industry was redefined to stop at the end of a wire on the customer's premises, and any equipment attached to that wire was excluded from the regulated industry. The boundary was defined with public interface standards. All equipment on the customer's side of the boundary was for the use of that customer only and was excluded from the complex revenue sharing and subsidy mechanisms of the regulated telephone industry.

The 1980 Computer II decision transformed customer premises equipment into a fully competitive market. It eliminated a great deal of price discrimination that had been imposed earlier by charging different prices for telephone service with different kinds of terminal equipment. For example, earlier data communication service was provided by a telephone company supplied modem attached to voice grade lines. It was offered in various speeds at different prices, with no direct relationship between the price of data communications service and the corresponding underlying price of a telephone line and a modem. After 1980, customers purchased simple terminal equipment on an open market and simply plugged it in with a publicly defined interface standard. More complex kinds of terminal equipment, such as a Private Branch Exchange (PBX), could also be supplied privately but required direct wiring into the network.

Opening the market to terminal equipment created a massive flood of new types of terminal equipment. The growing importance of data communications caused a great demand for modems, which were developed in great numbers largely outside the telephone companies. Ordinary telephones gained many new features and styles. Specialised terminal equipment became available and was only limited by the demand of customers, not by telephone company policies. The potential harm to the network from allowing non-telephone company equipment to be attached that was such a concern during the 1970's failed to materialise as competitive equipment proliferated during the 1980's. Although the 1980 decision was controversial at the time it was adopted, there was never any serious consideration, after it was implemented, of reversing the policy or of re-establishing the link between the telephone company network and customer terminal equipment.

3.2. Competition in long distance service

The 1948 restrictive microwave licensing decision effectively limited the market to AT&T. However, companies such as pipelines that needed communication in areas not served by AT&T were allowed to build microwave systems for their own private use. During the 1950's, a number of companies sought permission to build microwave systems for private usage. In 1959 the FCC largely reversed its restrictive 1948 policy in a decision known as "Above 890" because it referred to the utilisation of frequencies higher than 890 megahertz (Mhz). The highest television frequencies allocated for broadcast stations in the UHF band extended to 890 Mhz and frequencies higher than that were potentially available for microwave communication. In the late 1950's inquiry that led to the decision, the Electronic Industries Association submitted a detailed engineering study which showed that with proper engineering techniques adequate frequencies were available for both private and common carrier microwave usage. The Communications Act requires the Commission to: "Encourage the larger and more effective usage of radio in the public interest." Consequently, the Electronic Industries Association study was a critical piece of evidence. The Commission decided to license private users of microwave communication systems on a permanent basis and not require them to disband their facilities when common carrier capacity became available as in the previous decision.

Private microwave systems required a substantial fixed cost per route and a low marginal cost per circuit. The systems were not competitive for users with traffic spread over many different locations. However, they were competitive for companies that needed many private line circuits between two points. At the time of the decision, AT&T charged a fixed rate per circuit for private lines regardless of the traffic carried over that circuit and charged for multiple private lines at the multiple of the single line rate with no volume discounts. The cost of private microwave systems relative to the then existing tariff structure made common carrier service cheaper than private microwave for those with fewer than about fifty circuits between two specific points, but provided strong incentives for those with very high two point demand to build a private system.

The 1959 'Above 890' decision was a spectrum policy decision and not a determination that competition with AT&T was desirable. It authorised private users to build their own communication systems using microwave frequencies, but did not allow them to resell those circuits to others or to interconnect with AT&T. However, the decision was a critical link in developing long distance competition. Once large users could supply their own communication needs AT&T responded with a substantial volume discount plan (known as Telpak) which offered up to 85 percent discounts over the previous rate for bundles of 240 lines. That is, a very large user of private lines would pay fifteen dollars after the AT&T pricing response for service that had cost a hundred dollars prior to the response (Brock, 1994, pp. 105–108).

The cost of providing bulk private microwave communications or of receiving service from AT&T compared to single line rates created price discrimination between large and small users. In an ordinary competitive market entrepreneurs would arbitrage between those differences in order to profit in the short run and bring the prices together in the long run. Such arbitrage could occur either by building a microwave system and providing service to smaller users or by purchasing AT&T bulk rate Telpak service and reselling it to smaller users, but both of those actions were prohibited. AT&T prohibited the resale of its services and the FCC authorisation for private line microwave did not include the right to resell the service to others.

In 1962 a new company called Microwave Communications, Inc. (MCI) filed an application to build a public microwave system between St. Louis and Chicago. MCI proposed a system with limited capacity on a single route that could have been routinely approved if it had been filed as a private microwave system for MCI's internal use, but MCI intended to provide service to the public. It, therefore, needed a new policy authorisation. After extensive proceedings and vigorous opposition from AT&T and other established carriers, the FCC approved MCI's proposal on a four to three vote in 1969. It is unlikely that MCI's application would have been approved if MCI had simply proposed providing identical service to that provided by AT&T at lower cost. Although there was no explicit grant of monopoly to AT&T, the Commission had established a policy that duplication of facilities and services was generally unwise. Near the time of the initial MCI decision the Commission had denied a request from GTE to build its own microwave system to carry toll traffic among its telephone companies because the proposed system would duplicate AT&T facilities. However, MCI proposed 'specialised' services that would differ from the services provided by AT&T and would be customised to meet special needs. The differences between MCI's and AT&T's services were an important component in both the substantive reasoning and the legal justification for the decision.

After the approval of MCI's very narrow application to provide a limited set of specialised services between one city pair, a large number of applications for similar services were filed. Rather than examine each application individually, the Commission chose to establish general rules for specialised common carriers. In 1971 the Commission announced a general policy in favour of the entry of new carriers. The Commission concluded,

> "We further find and conclude that a general policy in favour of the entry of new carriers in the specialised communications field would serve the public interest, convenience, and necessity" (FCC, 1971, p. 920).

The 1971 decision clearly contemplated limits on the services provided by new carriers but did not specify precisely the meaning of specialised services. The Commission at that time was concerned with protecting the subsidy system in

which switched long distance rates were well above the cost of providing service, and revenues flowed back to the local exchange operating companies to subsidise local rates. Pressure from Congress and state regulatory commissions encouraged the Commission to maintain the subsidy system. The general understanding at the time was that the new carriers were limited to private line service and excluded from switched services.

In 1975 MCI introduced a service called Execunet. It described the Execunet service as a "shared foreign exchange service" and treated it as an extension of its existing private line service. Under the Execunet service, a customer dialled the MCI office as an ordinary local call and was connected to the 'private line' through the MCI switch. MCI's Execunet service was a version of the arrangement provided by AT&T in which an employee of a large corporation could call in to the company office and have the call transmitted along the company network to a terminating destination. The distinction was that in MCI's case anyone could participate in the network by subscribing to it and receiving an identification number that allowed calls to be placed. While MCI treated the Execunet calls as an extension of its private line service, they were in effect ordinary switched calls. The quality was lower than that provided by AT&T and the dialling sequence was longer but the arrangement allowed an MCI customer to reach any other customer.

AT&T objected to the arrangement and filed a formal complaint with the FCC. The FCC quickly ruled that the MCI Execunet tariff was beyond the scope of the MCI service authorisation. MCI appealed that order to the federal appeals court and won a major victory. The appeals court ruled that because the FCC had not specifically defined the scope of MCI's authorisation at the time that it was granted, it could not retroactively limit MCI's service authorisations for the existing facilities. It would have been legally possible for the FCC to undertake new proceedings in order to make a formal finding that competition in switched services was against the public interest. But it would have been difficult to justify that finding and such a finding would limit its future flexibility. Thus the appeals court ruling allowing MCI's service to continue effectively opened the switched market to long distance competition.

The court decision allowing MCI's Execunet service to proceed did not specify the rates to be charged for the local service portion. In MCI's initial service the local service portion was charged at ordinary local rates, typically with no charge per minute for residential customers and a very modest charge per minute or per call for business customers. Under that arrangement a caller initiated a long distance call by placing a local call to MCI's office. The call then proceeded across the MCI network, and the MCI switch placed a final local call to the terminating customer. That arrangement eliminated the subsidy payments that had been built into traditional long distance rates. AT&T objected vigorously to that pricing arrangement, and the FCC agreed with AT&T that MCI ought to pay at least a portion of the long distance subsidy paid by AT&T. However, computation of the

subsidy was complicated by the fact that it was administered by AT&T, and much of it consisted of internal payments between the AT&T long distance division and the AT&T owned local operating companies. MCI denied that any subsidy existed while AT&T claimed the subsidy was very large. The FCC worked out a temporary compromise in which MCI paid far higher than the local rates for its end charges but substantially less than what AT&T believed was the rate that it paid.

Both AT&T and MCI objected to the compromise rate. AT&T asserted that the compromise was far too low and that MCI was an inefficient company that could only compete because of favourable payments relative to those made by AT&T to the local companies. MCI asserted that it was a far more efficient company than AT&T but that its efficiencies were partially masked by high fees paid to AT&T.

MCI and a second competitor (Southern Pacific Communications, later folded into Sprint) grew rapidly from a small base in the early 1980's. MCI's gross revenues rose from $206 million in 1980 to $1.521 billion in 1983, an exponential growth rate of 67 percent per year for those three years. The companies initially built backbone networks among the major cities and then filled in connections to smaller cities. MCI was profitable in that early period. Between 1980 and 1983, MCI earned an average of 13.3 percent on total assets and 21.5 percent on stockholder equity. By 1983 the two primary competitors had achieved 6 percent of the total switched toll revenue reported to the FCC while AT&T retained 94 percent (Brock 1994, pp. 142–143).

After the AT&T divestiture in 1984 (Section 3.3), a system of access charges was implemented along with technically equal access for AT&T and the competitive carriers. Instead of competitive carriers receiving 'line side access' in which they were treated as an ordinary end user, carriers received a higher quality connection known as trunk side access. Furthermore, a system of pre-subscription was undertaken so that dialling '1' for long distance automatically took a subscriber to that subscriber's preferred long distance company. Along with those technical changes AT&T and its competitors began to pay the same charges for access. Initially those charges were over $0.17 per minute in combined originating and terminating access charges, and constituted approximately half of the total toll revenue collected by interstate carriers. The access charges were gradually reduced through a program of rate rebalancing, dropping to $0.095 in 1989, and to $0.019 at the end of 2000 as a result of many small changes (FCC, 2000, p. 1–6).

The movement to equal access and equal technical arrangements was initially very difficult for AT&T's competitors. Their financial situation declined as they were required to pay much higher charges for access to the local telephones than they had under the pre-divestiture compromise rates. There was considerable doubt during the mid 1980's about the financial viability of the competitors as a whole, and Sprint in particular, but the competitors adjusted to the new situation and continued to grow and to take market share from AT&T. AT&T's share of

the post-divestiture long distance market dropped from 90 percent in 1984 to 41 percent in 1999 while MCI's share rose from 5 percent in 1984 to 24 percent in 1999 for the combined MCI-WorldCom. The remainder of the market was served by Sprint and a large number of smaller long distance companies (FCC, 2000, p. 10–14).

The initial long distance competition was based on microwave technology and in particular on the ability of new entrants to quickly establish microwave routes without obtaining rights of way. By the mid 1980's, the technology of choice for high-density long distance communication began changing from microwave to fibre optic cables. The new carriers embraced the fibre optic technology quickly, and were able to arrange adequate rights of way to establish nationwide networks. In fact, the competitors' share of fibre optic capacity has been far greater than their share of total long distance revenues. In 1985 AT&T had 86 percent of toll service revenues but only had 28 percent of the total fibre system route miles reported by inter-exchange carriers. In that year, Sprint had almost as extensive a fibre system as AT&T even though it was a far smaller carrier in terms of revenue.

3.3. The divestiture

U.S. telecommunication policy is made by many different government agencies. The process is chaotic, with no central authority and with overlapping responsibilities. Conflicts among government bodies in telecommunication are routine. Common carriers are regulated both by the Federal Communications Commission and state regulatory commissions with murky lines dividing the respective jurisdictions. The FCC gains its authority from the Communications Act of 1934. However, that law did not explicitly exempt regulated companies from the antitrust laws. Consequently, there are potential conflicts between FCC decisions under its congressional mandate and Department of Justice decisions in its role as enforcer of the antitrust laws.

AT&T's aggressive actions against MCI's initial long distance competitive activity, and John DeButt's 1973 public challenge to the FCC's competitive policies, created concerns within the Department of Justice (DOJ) that AT&T was abusing its market position. Soon after the speech, the DOJ issued a Civil Investigative Demand as the first step toward an antitrust suit and the next year it filed an antitrust suit against AT&T. According to the DOJ, AT&T participated in three separate markets: local telephone service, long distance telephone service, and the provision of customer premises equipment or CPE. The local telephone service market was a natural monopoly subject to state regulation. The long distance and CPE markets were potentially competitive but were dependent upon interconnection with the local market. The Department of Justice asserted that AT&T used its legitimate market power in the natural monopoly local telephone service to illegally limit competition in the potentially competitive long distance and CPE markets.

The DOJ antitrust case was in part based upon the belief that the FCC was incapable of properly regulating AT&T, and that the FCC's regulatory policies actually aided AT&T in its effort to limit competition. The DOJ asserted that the FCC was not intentionally serving AT&T's interests but that it was overwhelmed by AT&T's abuse of process. The DOJ claimed that AT&T's regulatory interventions and court appeals went beyond the legitimate exercise of its right to plead its case, and constituted a pattern of engaging in litigation in order to raise the cost of rivals and prevent them from becoming active competitors of the Bell System.

AT&T's defence also put great weight on regulation but with a very different perspective than that of the DOJ. According to AT&T, it was subject to pervasive state and federal regulation that controlled prices and entry and limited AT&T's total profits. Although AT&T had a high market share it did not possess monopoly power in the legal sense (the power to control prices and exclude competitors) because regulators rather than AT&T controlled entry and prices. According to AT&T, the entire telephone business was a natural monopoly with economies of scale, scope, and vertical integration that would be lost if service were provided by multiple companies. AT&T acknowledged the facts of the particular incidents cited by the DOJ as evidence of anti-competitive intent. It, however, interpreted the incidents as the result of good faith efforts of a very large corporation to adjust its practices to confusing and contradictory policy changes by state and federal regulators. AT&T asserted that it did not engage in predatory pricing but sought to maintain pricing policies and revenue flows that had long been approved by regulatory authority and that it engaged in regulatory and legal activities as the legitimate exercise of its right to petition the government.

Initially, the suit moved slowly as AT&T unsuccessfully sought dismissal on the grounds that regulation precluded application of antitrust standards and while the parties exchanged documents and refined their positions. In 1978 the case was assigned to Judge Harold Greene who established a strict timetable for trial, requiring the parties to develop stipulations and detailed pre-trial contentions in order to begin the trial in early 1981.

Just before the trial began in 1981, Republican President Reagan replaced Democrat President Jimmy Carter and appointed a new set of top Executive Branch leaders. There was considerable dispute within the top levels of the new administration with regard to whether or not to continue the case and to pursue remedies that had been developed by staff members under the direction of the previous administration's political leadership. The Department of Defense opposed a divestiture and believed that a unified company could best supply national defence interests. The Department of Commerce opposed the divestiture because it believed divestiture would lead to greater foreign encroachments on the U.S. market for telecommunication equipment. However, William Baxter, the assistant attorney general for antitrust, shared the general perspective of DOJ economists that the problems in the industry came from a combination of

regulation and competition. He wanted to draw a clean line between the regulated monopoly parts of the industry and the potentially competitive parts of the industry. Baxter's general approach was consistent with the FCC's approach in the 1980 Computer II decision (Section 3.1), but Baxter wanted a stronger structural separation (divestiture rather than the FCC's separate subsidiaries), and he drew the lines around the monopoly parts of the industry more narrowly (including only the local operating companies rather than all basic network services) than the FCC had done. Baxter placed no confidence in regulation or injunctive relief. His goal was to remove the structural problems of the industry and then allow the competitive market to work without further government intervention.

The leaders of the respective cabinet agencies were not able to reach agreement on the most appropriate policy toward AT&T and the President declined to take an active role. Consequently, the Department of Justice continued the antitrust suit in accordance with its perspective of what ought to be done and its authority to enforce the antitrust laws. Simultaneously, the Department of Defense and the Department of Commerce helped sponsor an administration initiative in Congress to solve the competitive problems of the industry without a divestiture.

During 1981 the antitrust trial and the effort to craft a separate Congressional settlement proceeded simultaneously. By the end of that year, the DOJ had finished presenting its evidence and, before AT&T began its defence, it requested a directed verdict. Judge Greene delivered an extensive opinion reviewing the evidence in detail and showing that he was convinced by the Justice Department's presentation of the case against AT&T and that AT&T would have a difficult, if not impossible, job of rebutting the case. In late 1981, the legislative effort supported by the Departments of Defense and Commerce was overwhelming approved by the Senate but failed to pass in the House.

AT&T's options were limited, as it faced a decreasing possibility of a compromise legislative solution that would derail the antitrust suit and an increasing possibility of a guilty verdict if the trial were completed. Consequently, AT&T surrendered to the DOJ's divestiture request without presenting its defence and waiting for the entire legal process to play out. That approach gave AT&T more bargaining power in shaping the details of the divestiture and reduced the risk of adverse financial judgments from a guilty verdict in the courts. After brief negotiations at the end of 1981, the parties reached the settlement that became known as the Modified Final Judgment (MFJ) because its legal structure was a modification of an earlier antitrust settlement (the 1956 Consent Decree) between the DOJ and AT&T.

The goal of the MFJ was to leave the competitive or potentially competitive businesses with the new AT&T, and to separate those businesses from the monopoly local telephone companies that were part of the old AT&T. According to the divestiture theory, the local exchange carriers should be confined to the provision of monopoly local exchange service and prohibited from participating in the competitive markets of long distance service, manufacturing, or

information services. Participants in the competitive markets should be unregulated and unaffiliated with the LECs. They should purchase needed local services from the LECs on a non-discriminatory basis with prices for those monopolised inputs controlled by regulation.

The DOJ believed that such a division would remove both the incentive and the ability of the divested Bell Operating Companies to discriminate against competitors because they would have no incentive to restrict their access services to their own operations. AT&T would be dependent on the local exchange companies for access to its long distance and enhanced services, and for connection of its customer premises equipment, just as any of AT&T's competitors were dependent on those operating companies. In order to partially meet the concerns of the Department of Defense that divestiture would result in fragmented responsibility for responding to emergency communication needs, the MFJ required that the independent divested Bell Operating Companies provide a centralised organisation and single point of contact for national security and emergency communications.

Baxter believed that the brief decree provided a clear guide as to how the divestiture would occur. AT&T found the process far more complex than Baxter had envisioned. AT&T initiated a vast planning and implementation process to accomplish the divestiture. By the end of 1982, AT&T had transformed the general principles of the short MFJ into a 471 page plan of re-organisation. With over one million employees and nearly $150 billion in assets, AT&T's operations were complex and interrelated. Many difficult problems, ranging from assignment of personnel to division of jointly used assets to reprogramming operational systems to work independently, had to be solved. The divestiture process was completed over a two-year period of intensive work and became affective on January 1, 1984 (Temin and Galambos, 1987; Brock, 1994, pp. 149–172).

4. Efforts to develop interconnected competition

4.1. The competitive unregulated internet

As with many innovations, the Internet resulted from the efforts of multiple independent thinkers. In the early 1960's Paul Baran of Rand prepared theoretical papers on the design of a survivable communication system. His work suggested developing a network with redundant connections that had no hierarchy and was connected by computers with routing tables. Individual messages would be broken up into blocks that could be independently routed over the network and reassembled at the final destination. A second independent development resulted from the work of Donald Davies of the British National Physical Laboratory. Davies was concerned with the efficient transmission of data that arrived in bursts. Data communications often needs rapid transmission of a large burst of

data followed by dead time with no transmission. Davies recognised that the existing circuit switching structure did not handle that approach efficiently. He suggested a structure for data communication with many messages using a shared high capacity channel. Davies designated the units of a message as 'packets' using an analogy with small packages. His group built the first packet switching network on a small scale at the National Physical Laboratory. A third conceptual justification was developed by the U.S. Department of Defense Advanced Research Projects Agency (ARPA) which was concerned with allowing greater sharing of computer resources among various ARPA sites. ARPA proposed a plan to connect remote computers in 1967. After refinement from the efforts from Davies and Baran, along with Wesley Clark and numerous others, ARPA developed the first cross-country packet switching network.

The first node of the ARPANET was installed in 1969 at UCLA, by a team led by Leonard Kleinrock and including graduate students Vint Cerf, Steve Crocker and John Postell, who played major roles in the development of computer networking. The network had four nodes by the end of 1969. Graduate students from the first sites began communication with each other on developing ways to make the network useful. They developed many of the early protocols that were used in the later Internet. The approach utilised in the ARPANET was a departure from traditional communication or computer practices in which one large company sponsored a particular protocol or language. Instead, the graduate students formed a loosely organised network working group which communicated regularly with each other and distributed proposals known as "Request for Comments." The proposals eventually became standard parts of the Internet if they achieved widespread acceptance, but there was no formal procedure to adopt standards.

The ARPANET was a single centrally managed network. Each ARPANET site required a particular minicomputer running particular programs, and that computer provided the interface between the network and the host computers on that site. However, other packet switched networks were developed using different standards. During 1973 Vint Cerf and Bob Kahn (then at ARPA) developed a protocol for exchanging packets among networks operating on different standards. After further refinement, their paper was developed into the transmission control protocol (TCP, later refined into the TCP/IP protocol) that would allow data packets to be encapsulated in standard 'envelopes' that could be passed across networks. The TCP protocol was crucial for the later development of the Internet. It provided a standard way of carrying dissimilar packets. The TCP protocol also borrowed an idea from the French Cyclades network in which error correction was entirely the responsibility of the host computers. The network was assumed to be unreliable, and hosts were responsible for sending packets again if there was no acknowledgement of correct transmission. In contrast the ARPANET considered the network responsible for reliability, and built extensive acknowledgement and error correcting mechanisms into the message processors,

allowing the host computers to assume that messages had been correctly transmitted.

The change in error handling was a significant conceptual shift with far reaching consequences for the developing Internet. It recognised that the packets could be duplicated essentially for free, and that it might be more efficient to retransmit the messages rather than to ensure error free handling across the network. To use a shipping analogy, the ARPANET approach assumed that the contents of the containers were fragile and valuable. Therefore, it was worth extensive expense to ensure that the containers were handled carefully and delivered without damage. The TCP approach assumed that the contents of the containers could be replaced at a zero cost and that the only goal was to maximise the number of undamaged containers transmitted. That goal might be achieved by transmitting containers with little concern for rough handling or loss and simply replacing the ones that failed to arrive intact.

A second conceptual change in the TCP emphasis on error correction at the host was its general acceptance of locating the 'intelligence' of the system at the edges of the network under the control of individual users. The telephone network was a tightly controlled and centrally managed system. However, the emerging Internet was not centrally managed. Individual networks with varying standards and degrees of reliability could be attached to it and individual applications could be developed to add value to the network without approval of a central authority. The simplicity of the network created by the encapsulation procedure and the absence of guaranteed reliability provided a very open platform that was able to incorporate vast changes without being fundamentally redesigned. The absence of formal standards, reliability requirements, and central control also made the early Internet appear to professional telephone engineers to be inferior to the stable, standardised, carefully managed telephone network.

In 1985 the U.S. National Science Foundation (NSF) built a backbone system to connect five supercomputer centres, and to allow academic institutions to build regional networks and interconnect them to NSF backbone. As the new interconnected networks proliferated, the original ARPANET became increasingly obsolescent and was shut down in 1989. The very open nature of the Internet allowed commercial networks to utilise the TCP/IP protocol and interconnect freely. During 1991 commercial networks began interconnecting with each other to bypass the NSF backbone, and avoid restrictions on commercial usage put forth by the National Science Foundation. With the rapid growth of commercial networks, fuelled in part by the development of the World Wide Web and optimistic financial projections of the commercial potential of the Internet, the NSF backbone was shut down and the Internet became an interconnected set of privately funded commercial networks (Hafner and Lyon, 1998; Abbate, 1999).

The primary U.S. regulatory framework for the Internet is the FCC's 1980 Computer II decision. That decision was based on an examination of the interaction between the computer and the communication industries and was not

primarily designed to regulate the Internet. At the time of the decision, the ARPANET and other individual networks existed and some were beginning to interconnect using the TCP/IP protocol but the Internet as such hardly existed. All of the networks were small and formed an extraordinarily tiny fraction of the communications capacity. As discussed in Section 3.1, the Computer II decision completed the long controversy over the regulatory treatment of customer premises equipment (CPE) by deregulating it entirely. The CPE portion of the Computer II decision created a right for customers to provide services within their own premises and then to connect those services to the public switched network. Without that decision, the task of connecting an internal local area network to the wider network would have been legally suspect or subject to arbitrary price increases by the telephone companies. The Computer II decision eliminated the incentive and ability of the telephone companies to dominate the emerging data communications market through their control of network interfaces and interconnection conditions.

The CPE decision also allowed and encouraged the development of intelligence at the customer's premises. Sophisticated individual customers developed complex internal systems (both voice and data) without telephone company control. The Internet concentrates the intelligence of the network at the edges in individual computers and local area networks while the telephone network keeps the intelligence within the network in a set of complex multifunction switches. Internet routers can be much simpler than telephone company switches because they perform far fewer functions. Adding new applications to the voice network requires upgrading the central office switches with new hardware and/or software. Adding new applications to the Internet only requires upgrading the capabilities of an application server at the edge of the network.

The 1980 Computer II decision also created a new legal category called enhanced services. The enhanced services category was the result of a long controversy over the boundary between regulated communication service and unregulated data processing service. The essential distinction was that enhanced services used computer power to add additional functions to basic service. It was assumed that an enhanced service provider was a reseller that procured basic telephone service from a regulated common carrier, and then added computer processing to enhance the service and make it more suitable for whatever specialised applications the provider targeted. The FCC declared that enhanced services were not common carrier communication services subject to the regulations imposed on telephone companies, but also asserted that the FCC had full authority over enhanced services and would prohibit the states from regulating them. That legal construct created an opportunity for an unregulated communications service. It gave great freedom to individuals to provide enhanced communication services without filing tariffs or meeting other regulatory requirements. It also guaranteed service providers the right to utilise the existing common carrier services in any combination with their own enhancements.

The FCC's enhanced services category provided a favourable regulatory environment for the development of the Internet. While data communication was correctly forecasted to grow rapidly, there was not enough demand to economically utilise a separate data communications infrastructure during the early Internet period. The Computer II enhanced services category allowed the developing Internet to freely utilise the extensive common carrier communications network while also exempting the Internet from the regulatory burdens associated with common carrier network.

The favourable regulatory framework for the Internet allowed it to escape the established practice of penalising new technology in favour of old technology. With many communication innovations, the regulatory pricing practice imposed a relatively high price on new services and used the resulting revenue to subsidise older technology. The justification for this practice was a social policy in which it was assumed that higher income consumers utilised the newer technology and lower income consumers utilised the older technology. The regulatory pricing practices were a form of price discrimination that slowed the spread of new technology and services in the telephone network. Had that practice been applied to the Internet, it is likely that data communications would have been priced at a relatively high rate and used to subsidise the other services. However, the Internet regulatory framework prevented regulators from extracting high fees from Internet users.

The Computer II regulatory advantages of the Internet were increased by two later decisions related to universal service. The first was the access charge exemption for enhanced service providers. When access charges were developed during 1983 (to be implemented with the divestiture), a number of special categories were exempted from the access charge arrangements because of special administrative difficulties. Those exemptions were not meant to be long-term policy but simply deferments until the special conditions related to particular services could be evaluated and appropriate rules put in place. In most cases, the exemptions were removed. However, when the FCC proposed eliminating the access charge exemption for enhanced service providers (including Internet access providers) in 1987 a major political confrontation ensued. At that time the small but rapidly growing Internet had enough constituents and users to create political pressure for maintaining the exemption, but it was still a very small part of the overall communication structure. If access charges had been imposed on enhanced service providers at that time it would not have created enough revenue to make any significant difference in the overall level of access charges. Consequently, it was easy for supporters of the exemption to argue that Internet access was different from ordinary long distance access and ought to be protected and that, furthermore, no one was being seriously hurt by the exemption. The exemption was therefore continued and has remained in place to date.

The access charge exemption allows Internet service providers to offer flat rated service because subscribers can use ordinary local lines to reach them at ordinary

local rates rather than the higher interstate access rates that are imposed on long distance carriers. In effect, the access charge exemption allows Internet service providers to carry out the originally planned Execunet service that MCI proposed during the 1970s in which a caller simply makes a local call to the long distance provider and then goes out over the network. The access charge exemption provided a substantial benefit to Internet Service Providers compared to voice long distance providers when access charges were high, but the differential treatment has been greatly narrowed as access charges for long distance providers have dropped from $0.17 to $0.02 per minute.

4.2. The Telecommunications Act of 1996 and local telephone competition

From the time the divestiture was implemented, the Bell Operating Companies began seeking removal of the restrictions on their activities. The manufacturing, long distance, and information services prohibitions appeared to them to limit their flexibility and opportunities for technological progress. The divestiture theory treated local exchange as a mature well defined technology. The policy problem was simply to avoid the misuse of market power, not to promote technological change. The restrictions were designed to force the Bell Operating Companies to focus their business plans on providing a high quality product at low cost. The Bell companies vigorously disagreed with the restrictions and sought much greater freedom to develop new products and expand into new markets. They pursued their efforts to remove the restrictions through petitions for modification of the decree before Judge Greene (who administered the detailed interpretation of the agreement), through efforts with the FCC, through efforts with the Executive Branch, and through efforts to get Congress to pass a new law that would grant them greater freedom.

After extensive proceedings before the FCC and Judge Greene the information services restriction was lifted in 1991, but the other two restrictions remained. Judge Greene lifted the information services restriction reluctantly, as a direct result of an order from the appeals court, and indicated he was unlikely to remove or relax the other restrictions. The Bell companies focused extensive lobbying efforts on Congress in an effort to get a new law that would overturn the MFJ restrictions. In 1994 and 1995, progress towards a new law was made by rejecting the natural monopoly assumption of the MFJ. The new policy formulation assumed that local exchange telephone service was not a natural monopoly but was monopolised by the efforts of the incumbent telephone companies and the state regulators. The new policy approach allowed a political trade-off in which the Bell operating companies would give up their monopoly position in local exchange in return for freedom to enter the long distance and manufacturing markets.

After extensive efforts, a new law was passed in early 1996. The Telecommunications Act of 1996 was structured as a major amendment to the

Communications Act of 1934, not as an entirely new law. Consequently, it preserved many features of the existing policy structure including the FCC and all of the then-current regulations made by the FCC. The 1996 Act specifically abolished the MFJ and replaced it with a list of conditions that the Bell Operating Companies needed to meet in order to gain the legal right to enter long distance service. Under the new law, the decision to allow a Bell Operating Company to initiate long distance service from its territory was made by the FCC after an evaluation of the company's compliance with a statutory list of market-opening conditions.

In return for increased freedom to enter new markets, the new law required the incumbent local exchange carriers to open their monopolised markets to competition. It required local operating companies to interconnect with new competitors on a reciprocal compensation basis. The law also required incumbent local exchange carriers to provide unbundled network elements, such as basic copper wires, to competitors and to provide their services on an wholesale basis so that competitors could obtain those services and distribute them under their own brand name. The new law explicitly prohibited state legal restrictions on competition.

The local competition provisions of the 1996 Act were based in part on the earlier efforts of a few state regulatory commissions to develop competitive policies for local exchange service. Local telephone service had been considered a state matter and states varied on their approaches to competition. Many states prohibited competition altogether but some sought ways to allow competitors to develop in the local market. Some of the practices developed in those state commissions were incorporated into the 1996 Act. That Act was a federal law that applied to the entire country but it left considerable authority to the individual states. The law required companies seeking to enter a market to first enter into negotiations with the incumbent company and attempt to reach a voluntary agreement. It also specified that the state regulatory commission would develop a mandatory arbitration arrangement if the parties did not reach a voluntary agreement.

The 1996 Act was a direct repudiation of the MFJ itself, and of the theory behind the MFJ. Rather than carving the industry into legally defined segments, and limiting movement among those segments, the 1996 Act assumed that competition was possible in all parts of the market. The empirical support for that proposition was weak at the time of the Act. The increasing availability, however, of local fibre optics rings connecting business buildings, of wireless communication services, and of communications over cable television networks have made it possible that substantial competition to the telephone company copper wire network could develop.

The implementation of the Telecommunications Act of 1996 has been slow and difficult. The new law did not clearly specify which parts of the Communications Act of 1934 were being modified and which parts remained the same, leaving

ambiguity and even contradictions in the text of the amended law. In particular, the new law modified the previous division of authority between the state and federal regulatory agencies but did not amend the old provisions regarding the division of authority. The ambiguity of the new law created extensive litigation, and even after a Supreme Court decision legal issues related to the new law remained unsettled after five years of implementation efforts. The Bell Operating Companies believed that the FCC's orders interpreting the local competition provisions of the new law were too favourable to competitors and fought those provisions through the courts. The FCC believed that the Bell companies' efforts to comply with the market opening requirements were inadequate and refused to grant permission for them to initiate long distance service from their territory. Five years after the 1996 Act was passed, Bell companies have been authorised to provide originating long distance service in only three states.

Competition in the previously monopolised local exchange markets developed slowly. In the year before the Act was passed, FCC data showed a total of 57 companies participating in some way as competitors to the local market with a total market share by revenues of 0.7 percent. Many of those companies were competitive access providers that focused on carrying traffic between large business customers and the inter-exchange carrier points of presence. That service did not require interconnection with the local exchange companies and was treated as jurisdictionally interstate, allowing it to occur even without the permission of the state regulatory commission. Many companies entered immediately following the Act, and between 1995 and 1999 the number of competitors increased from 57 to 349 while the market share by revenues increased from 0.7 percent to 4.1 percent (FCC, 2000, p. 9–8). The numbers for competitors and market shares do not include wireless providers.

A separate source of local competition was the tremendous growth in cellular and Personal Communication System (PCS) telephone subscribers. Cellular systems in the United States began operation in 1984 on a duopoly basis in which one license in each market was reserved for the local telephone company and the other was open to competitors. The cellular licenses granted 25 megahertz to each licensee in a narrowly defined geographical area. Later, substantial new spectrum was allocated to a second generation PCS service with the voice band portion of that spectrum divided into six separate blocks of differing sizes that totalled 120 megahertz. That spectrum was auctioned beginning in late 1994 and services were established soon afterward using digital technology. In June 1995, there were 28 million wireless subscribers in the U.S. and five years later that number had grown to 97 million (FCC, 2000, p. 12–4). While it remains rare in the U.S. for a person to use a cellular phone as a complete substitute for the landline phone, the extensive competition among numerous wireless providers, and continued technological progress, have made wireless service an effective competitor to incumbent telephone companies.

The Telecommunications Act of 1996 was not specifically directed toward either of the two most effective forms of competition to the incumbent local

exchange carriers: Internet access and wireless phones. Yet despite the backward looking nature of the Act (dealing with problems left over from the 1984 divestiture rather than wireless or Internet issues), and its slow contentious implementation, the 1996 Act marked an important change in policy perspective. For the first time since the regulated period began in the early 20th century, the U.S. national policy framework for telecommunication is based on competition in all parts of the market. The long standing concern with defining the boundaries of the monopoly, and controlling interconnection conditions between the monopoly and competitive services, was replaced by an effort to develop interconnected competition in all parts of the market.

5. Conclusion

During the late 19th century, the U.S. telephone and telegraph companies were a fast and expensive private alternative to the government owned and operated Post Office. The low Post Office rates provided a limit on the market power of the telephone and telegraph companies even when there was an unregulated monopoly provider. As telephone competition increased at the beginning of the 20th century, AT&T successfully sought a regulated monopoly, allowing political control of its prices in exchange for political protection from competition. The regulated monopoly telephone industry remained stable until the late 1960's.

Over a thirty year period (from the 1960's to the 1990's), the regulated monopoly structure was gradually transformed into a competitive structure. There was no single political or technological change that created the transformation. The transformation was initiated by the rapid technological progress in computers and the difficult regulatory problem of defining a boundary between regulated communication and unregulated data processing. The early concerns with providing 'specialised' equipment and services to meet needs poorly served by AT&T's generic services gradually evolved into a policy of allowing limited competition. The very open U.S. policy process, with power diffused among multiple institutions, made it difficult to limit competitors to specific niche roles. Once new companies were allowed to participate in the market, they became effective advocates for expanding their role. Generally, the FCC, DOJ and actual or potential entrants supported the expansion of competitive opportunities while the state regulatory commissions, AT&T, and the independent telephone companies supported a narrow interpretation of allowable competition. Once the end-to-end monopoly responsibility of the telephone company was abandoned, no stable dividing line between the monopoly and competitive portions of the industry could be found. A number of boundary lines were proposed and the most durable one was created by the AT&T divestiture, but all of the boundaries eventually gave way to the combined forces of changing regulatory policies and advancing technology.

The Telecommunication Act of 1996 abandoned the long effort to create a durable monopoly boundary and substituted an assumption of interconnected competition in all parts of the network. Direct competition with wireline telephones has developed slowly, and the incumbent telephone companies remain the only source of traditional telephone service for most customers. However, the rapid growth of wireless phones and the Internet have provided indirect competition to the incumbent telephone companies and have limited their market power. It appears likely that the importance of regulation will continue to decrease and that the telecommunication industry will continue to evolve toward a free market model rather than the past political control of the industry.

References

Abbate, J., 1999, Inventing the Internet, Cambridge, MA: MIT Press.
Appeals Court, 1956, Hush-A-Phone Corporation v. U.S. and FCC, 238 F. 2d 266.
Appeals Court, 1976, North Carolina Utilities Commission v. FCC, 537 F. 2d 787.
Brock, G., 1981, The telecommunications industry: The dynamics of market structure, Cambridge, MA: Harvard University Press.
Brock, G., 1994, Telecommunication policy for the information age: From monopoly to competition, Cambridge, MA: Harvard University Press.
Buderi, R., 1997, The invention that changed the world: How a small group of radar pioneers won the Second World War and launched a technological revolution, New York: Simon and Schuster.
FCC (Federal Communications Commission), 1939, Investigation of the telephone industry in the United States (Washington D.C, U.S. Government Printing Office.), reprinted Arno Press, 1974.
FCC, 1971, Federal Communication Commission reports, 2nd Series, 29, 870–920.
FCC, 1980, Federal Communication Commission reports, 2nd Series, 77, 384–495.
FCC, 2000, Trends in telephone service, Industry Analysis Division, Common Carrier Bureau.
Hafner, K., and M. Lyon, 1998, Where wizards stay up late: The origins of the Internet, New York: Simon and Schuster.
Henck, F., and B. Strassburg, 1988, A slippery slope: The long road to the breakup of AT&T, Westport, CT: Greenwood Press.
Holcombe, A., 1911, Public ownership of telephones on the continent of Europe, Cambridge, MA: Harvard University Press.
John, R., 1995, Spreading the news: The American postal system from Franklin to Morse, Cambridge, MA: Harvard University Press.
Temin, P., with L. Galambos, 1987, The fall of the Bell System, Cambridge: Cambridge University Press.
U.S. Bureau of the Census, 1975, Historical statistics of the United States, Colonial Times to 1970, Bicentennial Edition, Washington, D.C.: U.S. Government Printing Office.
Vail, T., 1915, Some truths and some conclusions, speech to Vermont State Grange, December 14, 1915, AT&T Historical File.
Weiman, D., and, R. Levin, 1994, Preying for monopoly? The case of Southern Bell Telephone Company, 1894–1912, Journal of Political Economy, 102, 103–126.

Chapter 3

NETWORK EFFECTS

STANLEY J. LIEBOWITZ

University of Texas at Dallas

STEPHEN E. MARGOLIS

North Carolina State University

Contents

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Network externalities

Network externalities, more properly called network effects, are modern versions of concepts that have existed in the telecommunication literature for decades, and in the economics literature for well over a century[1]. The basic idea is simple. As the number of users of a product or network increases, the value of the product or network to the other users changes.

Harvey Leibenstein (1950) undertook a general and early discussion of markets with these characteristics. He defined 'bandwagon' effects as occurring when individuals demand more of a product at a particular price when a larger number of other individuals demand the product and 'snob' effects as the reverse of the bandwagon effect. A third effect, called a 'Veblen' effect was defined as occurring when the utility derive from a product depends positively on its market price (not necessarily the price paid by the consumer). His discussion, particularly of bandwagon effects, is an early precursor for much of the current discussion of network effects, even though the causes of the network effects in which he was interested tended to be more sociological in nature. The market consequences of these characteristics that were of interest to Leibenstein, i.e. the impacts on the slope of the market demand, are quite different from those that interest modern network-effect theorists.

Leibenstein's main result was that demand curves are more elastic when consumers derive positive value from increases in the size of the market, that is, where there are bandwagon effects. His analysis, following the then traditional analysis of external effects, was focused on determining the market equilibrium, not on the welfare impacts due to 'network effects.' His focus might have been due to the fact that the external effect literature was mainly concerned with pecuniary externalities which, as we explain below, do not have any welfare implications.

Concepts such as this bandwagon effect are of particular importance in telecommunication industries. Models of telephone systems, for example, must deal with the empirical reality that the values consumers derive from a telephone system is a function of who and how many individuals can be contacted through the network (Rohlfs, 1974). In some of these models (Squire, 1973), the utility that consumers receive is not a just a function of the size of the network *per se*, but also depended on how many calls the network allowed the user to make or receive. In other models, such as Artle and Averous (1973) it is assumed that only the size of the network counts, an assumption adopted in the current network effects literature.

Squire hypothesized a 'conceptual' demand curve drawn for different size networks and constructed in a manner almost identical to that used by Leibenstein. Squire then determined an optimal pricing system, one that internalizes the external effects that occur as the number of subscribers changes. As is common in

[1] Leibenstein refers to several early works on these types of topics, with the earliest being Rae (1834).

traditional externality models, the externality can be internalized through creative pricing. Interestingly, he concludes that deficits likely to arise from his optimal pricing system might be made up by charging the individual receiving the call (as well as the person making the call) an arrangement that is now taken for granted among many users of cellular phones. He and other analysts of his time did not contemplate the possibility that external effects may cause the network to be the 'wrong' network, the focus of the modern network effects literature.

The modern network literature defines network effect as a change in the benefit, or surplus, that an agent derives from a good when the number of other agents consuming the same kind of good changes. This is really no different from the definition in the earlier literature. In many writings in this modern literature, particularly the early writings, all such effects were termed externalities rather than effects (Katz and Shapiro, 1985), implying a market failure.

Where the value of a good to consumers changes with the number of users, that value can be separated into two distinct components. One component, which elsewhere we have labeled the autarky value (labeled 'functional demand' by Leibenstein), is the value generated by the product if there are no other users. The second component, which we have called synchronization value, is the additional value derived from being able to interact with other users of the product. It is this latter value that is the essence of network effects.

Network effects, in principle, can be positive or negative. Congestion, a negative network effect, is created by the negative impact of additional users on network performance as capacity constraints are approached. The almost exclusive focus of the modern network effect literature, however, has been on positive impacts.

2. Classifications

There are three important classifications that arise in this literature: effects versus externalities, literal versus virtual networks, and direct and indirect effects.

Network effects are not usefully called network externalities unless the participants in the market fail to internalize these effects. Fully internalized economic interactions are not usually called externalities. A shopper, for example, imposes costs on a retailer, although the costs are fully compensated by the price paid for the goods purchased. It would make the term externality considerably less useful if we nevertheless took to referring to that interaction as the shopping externality. Externality has a specific meaning within economics, one that connotes market failure[2]. The use of the term 'externality' to mean something different in this

[2] One non-network externality that exists in the traditional telephone market is that the individual receiving the call does not pay for it, as noted in Squire (1973). Such is not the case for mobile, or cellular phones, however.

literature than it does in the rest of economics is likely to create confusion. Unfortunately the term externality has sometimes been used carelessly in the networks literature[3].

Although individual network participants are not likely to internalize the effects that their joining a network has on the network's other participants, the owner of the network may very well internalize such effects. When the owner of a network (or technology) is able to internalize such network effects, they are no longer externalities. Interestingly, this point was recognized by the earlier writers who were dealing with the real and specific case of telecommunications (Littlechild, 1975, p 667), but it was initially overlooked in the early modern treatments of network effects. Liebowitz and Margolis (1994) argue that network owners can be expected to internalize these effects, so that network effects need not be externalities, a point that now seems to be acknowledged many authors (e.g. Katz and Shapiro, 1994) but is not yet universal.

Literal networks are those that can be characterized by some means of connecting people to one another. The telephone system is clearly a literal network. Accordingly, the telecommunications literature focuses on literal networks. The Internet, a means of connecting people, is a literal network. In contrast, virtual networks are groups of consumers who are not literally connected by any facility, but who nevertheless place a value on the number of others who use the product. Liebenstein's examples of consumers who value products because of their popularity, or rarity, are examples of virtual networks since there is no direct physical linkage between these consumers. So too is the network of WordPerfect users, who can interchange files and expertise with each other, and therefore benefit from the presence of other WordPerfect users without actually being connected. This type of virtual network is often based on compatibility, an issue that reaches great prominence in software and electronics markets. But the concept of virtual network can be used to describe a much wider set of markets, such as the market for automobiles, where the ease of repair is a function of the popularity of a car and the inventory of parts that such popularity engenders. In some sense, most any market can be understood as a virtual network, although for many products network effects would be extraordinarily small.

Economists have discussed two types of network effects, direct and indirect, a distinction due to Katz and Shapiro (1994). Direct network effects have been defined as those generated through a direct effect of the number of purchasers on the value derived from a product, e.g. fax machines. Indirect network effects are "market mediated effects" such as cases where complementary goods, e.g. toner cartridges, are more readily available or lower in price as the number of users of a good, laser printers, increases. Note that direct effects can exist in either literal or virtual networks. For example, the number of AOL instant messenger users affects the value of the service and is a direct impact on a literal network. The

[3] See also Economides (1996) for a treatment of the modern networks literature.

number of WordPerfect users, on the other hand, is a direct impact, for file transfers, at least, on a virtual network.

Liebowitz and Margolis (1994) demonstrated that the two types of effects have different economic implications, a demonstration that relied upon resolutions of problems that had occupied economists in the early part of the twentieth century. Indirect network effects are, in most instances, pecuniary in nature, which is to say the interaction among individuals is manifest through effects on price. Pecuniary externalities do not impose deadweight losses if left uninternalized, whereas they do impose monopoly or monopsony losses if internalized. This resolution of pecuniary externalities goes back to Young (1913), and Knight (1924), and is fully exposited in Ellis and Fellner (1943). The early modern literature on network externality discusses direct and indirect network externalities, but does not acknowledge that the two have distinct economic properties. As a consequence, this literature repeats some mistakes of Marshall, Pigou, and others regarding pecuniary externality. It is now generally agreed, however, (Katz and Shapiro, 1994) that the consequences of internalizing direct and indirect network effects may be quite different.

3. Impacts of network effects

There is an important difference in focus between the recent literature on network effects and the earlier literature of telecommunications networks. This older literature focused on the traditional concerns in markets with externalities: would the network be the correct size to maximize total surplus, or social welfare? Arguments about universality of service, or related policy measures like common carrier obligations or lifeline service requirements reflect this economic concern with network size. A related question is whether unregulated network pricing would lead to the efficient size of a network.

In contrast, concern about marginal adjustment of the level of network activity has not been the primary focus of modern network externality modeling. Instead, this newer network literature has focussed on a different set of problems, including the selection of a winner among competing networks (Katz and Shapiro, 1985; Farrell and Saloner, 1985; Arthur, 1996). An important claim in this literature is that markets may adopt an inferior product or network in the place of some superior alternative. Choosing the wrong network would seem to dwarf problems of the wrong size network.

Farrell and Saloner use the terms 'excess inertia' and 'excess momentum' to describe situations where a lack of coordination causes consumers either to fail to adopt a better product, or, in the case of excess momentum, to adopt a new technology when it would be more efficient to remain with the old. Simply put, excess inertia would occur if market participants were unwilling to move to a new, superior, technology because of a fear that they would lose important network

effects if others in the market do not abandon the old technology. Excess momentum would occur if a group of users switched to a new technology they prefer without taking into account the costs imposed on a large group of other users who prefer the old technology. Network effects then induce these latter users to switch technologies even if the latter group is worse off than had everyone remained with the old technology. Excess momentum occurs if this loss is larger than the gain to the first group.

Much of the attention to the literature of network effects is owed to the view that such effects are endemic to new, high-tech industries. Thus the claim that markets are unable to chose efficient networks suggests that markets may be inadequate for managing new technologies.

3.1. Choosing network size

The level of network activity has received fairly little attention in contemporary discussions of network externality, perhaps because it is well handled by the earlier literature on networks. But network size is a real and significant issue that is raised by the presence of network effects.

Each agent who joins a network captures a benefit of participation, even as each agent also confers a possibly unremunerated benefit on all of the other network users. This simple reasoning has let some people to call these effects externalities. Where a network is not owned or 'sponsored' (Katz and Shapiro, 1986), and where there are no transactions among the users to internalize network effects, then the inefficiency that is implied by the term externality could occur. In such a case, where the network effect is positive, the equilibrium network size would be smaller than is efficient. For example, if the network of telephone users were not owned, it would likely be smaller than optimal since no agent would capture the benefits that an additional member of the network would confer on other members. This might also be thought to apply to a network of many compatible nodes or small networks, each individually owned, a view sometimes taken of the way the Internet might evolve. Symmetrically, a congestion effect, which is a kind of negative network effect, would imply that networks tend to be larger than optimal. If bandwidth were scarce, at least at certain peak times, this type of externality might plague many networks, including the Internet, (for pricing mechanisms to internalize these effects see Mackie-Mason and Varian, 1995). Again, this would constitute a true externality. Where networks are owned, however, things change.

The network problem is actually a mirror image of the open access, common property resource problem, what is known as the fisheries problem, or the 'tragedy of the commons.' To be more precise, positive network effects are the mirror image of the fisheries problem. The fisheries problem itself is a negative network effect. Fishermen participate in a network, with each additional

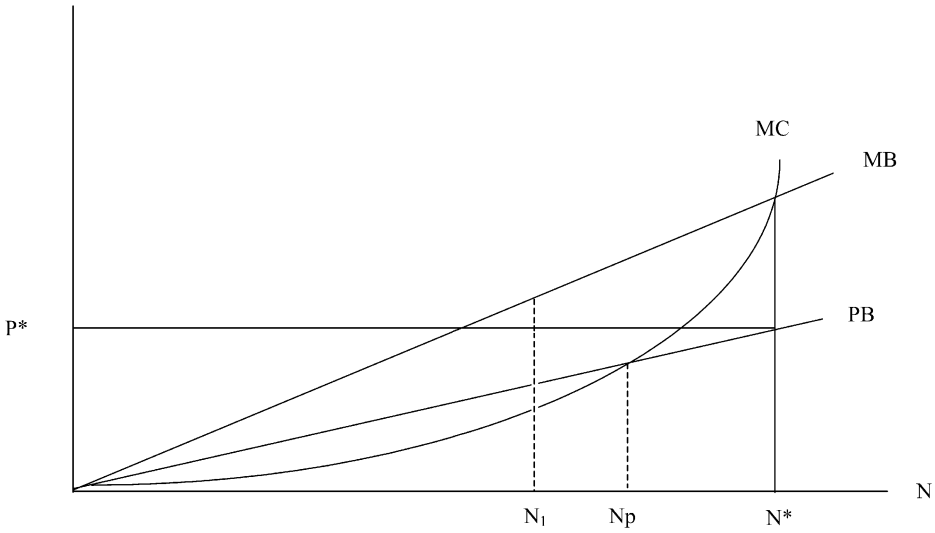


Fig. 1. Network size.

fisherman decreasing the value of network participation for all of the others. The following comes from Liebowitz and Margolis (1995b).

In the usual common property resource problem, the average benefit to a resource user declines as the number of users increase. This condition implies that the marginal benefit created by an additional user is less that the average benefit. The well-known inefficiency here is that each user enjoys the average benefit, so users are drawn to the common property so long as the average benefit exceeds their opportunity cost of using the resource. This means that in the open access equilibrium, the marginal benefit created by an additional user is less than the marginal user's opportunity cost. Absent any control over access, the resource, one could say the network, is exploited to the point that it confers to each user only their opportunity cost, so that the resource yields no surplus to the people who own the resource in common.

As is well known, profit-maximizing behavior by a private owner solves the problem for the basic case. A private owner of a fishery will hire the number of fishermen that equate marginal benefit with their opportunity cost, or will set a price for use of the resource that brings about the same level of use. In doing so, the owner would capture the maximum benefit from the resource.

The opposite case is positive network effect. Figure 1 illustrates this case. Positive network effects is the case that applies to most telecommunications networks, more communicators is better, though even telecommunications can be subject to crowding.

In the diagram, N is the number of users. The line PB is the private benefit of network participation that any user enjoys and is the maximum that any user would be willing to pay for access. Its upward slope reflects the presence of positive network effects. The line labeled MB is the marginal benefit of an additional user. Defining TB to be total benefits, then:

$$TB = N * PB \quad \text{and} \quad MB = \frac{\mathrm{d}TB}{\mathrm{d}N} = PB + N\frac{\mathrm{d}PB}{\mathrm{d}N}.$$

For the case of positive network effects, MB is greater than PB because the private benefit of network participation increases with N.

MC, the marginal cost of serving network participants, is shown as upward sloping in the diagram. Increasing marginal cost reflects the characteristic of many telecommunications networks that increases in participation requires connection customers who are increasingly difficult to connect. This would be the case, for example, for a cable network that must serve more distant customers, or less dense areas, as the network expands. Kahn (1988, Vol. 1 and Vol. 2), dealing with various conventional public utilities, argues that marginal cost with respect to network size increases as networks get large because marginal customers eventually are those that are more distant or otherwise more difficult to serve. Such concerns may well have their counterparts in the world of notional networks. Wireless networks require antennas or other facilities in very remote or low density areas as network size approaches its maximum possible size. An upward-sloping marginal cost function is not essential, however, and we consider the alternative below.

The optimal network size is shown as N*, where marginal benefit of network participation is equal to marginal cost. There is the possibility of a market failure here, as would be the case if private ownership were highly fragmented. For example, a communications network in which each participant owned privately his own transmitter-receiver, might exhibit this problem, unless there were some other mechanism for internalizing the network effects. Such a case is roughly analogous to the open access resource, and could be thought of as an open access network. In such a case, if each enjoys benefits PB and faces costs MC, the equilibrium network size will be N_P, which is less than the optimal network size.

The network model illustrated in Figure 1 parallels the fisheries model in most respects. The network model assumes that the potential network participants are homogeneous, as does the simple version of the fisheries model. Absent that assumption, monopoly pricing, including price discrimination, could be brought into either model. Profit maximization by a single private owner in the fisheries model brings about efficient use of the resource. But that result is obtained only in the absence of a monopoly in the fish market and monopsony in the labor market, so the standard model can hardly be said to be a general proof that there are no welfare

losses under private ownership. Nevertheless, the standard fisheries model provides a useful base case that offers important teachings.

The simple model of a positive network effect, as illustrated in Figure 1, offers similar lessons. One can assume alternative ownership structures, or heterogeneous participants, or other constructions of technology, in which private ownership will not bring about optimal network size. Nevertheless, Figure 1 offers a useful base case.

Figure 1 is constructed with marginal cost increasing sufficiently rapidly with the number of network participants that cost is the factor that limits network size. However, the analysis goes through even where the number of potential participants is smaller than N*. If N_1 in Figure 1 were the number of potential participants, then N_1 is also the optimal network size. The network owner would charge a price equal to the private (average) benefit at N_1. Marginal cost and marginal benefit would not be brought into equality, and the network owner would earn rents even at the margin. Nevertheless the network size, consisting of all potential users, is optimal.

Perhaps surprisingly, the problem of internalizing the network externality is largely unrelated to the problem of choice between competing networks that is taken up in the next section. In the case of positive network effects that are not internalized, networks are too small. In such cases, it is not the relative market shares of two competing formats but rather the overall level of network activity that will be affected by this difference between private and social values. This is completely compatible with standard results on conventional externalities.

3.2. Choosing among competing networks

Network effects are closely related to the economics of standards, where a primary concern is the choice of a correct standard (Farrell and Saloner, 1985; Katz and Shapiro, 1985; Besen and Farrell, 1994; Liebowitz and Margolis, 1996) and technology choice (Arthur 1989). These choices exhibit network effects. A standard is generally more useful as more people use it. Technology users may benefit from the development experiences of other users. Network effects have also played a role in discussions of path dependence (Arthur, 1989, 1990; David, 1985; Liebowitz and Margolis, 1990, 1995a).

Positive network effects are a kind of scale economy. Positive network effects, which raise the value received by consumers as markets get larger, have impacts that are very similar to conventional firm-level economies of scale. Under the assumption that different firms produce similar but incompatible products (networks), and that the network effects operate only within the class of compatible products, then competitors (networks) with larger market shares will have an advantage over smaller competitors, *ceteris paribus*. Where larger networks have a forever widening advantage over smaller ones, the network effect

creates a natural monopoly. Much of the literature on network effects or externalities do in fact postulate conditions that amount to natural monopoly.

It is important to note, however, that network effects are not generally sufficient for natural-monopoly-type results. Many, if not most, models in this area ignore production costs, and focus entirely on the scale economies that originate on the consumer or user side of the market. Thus the assumption of any positive network effects leads these models to natural monopoly results. In cases where average production costs are falling, constant, or nonexistent, network effects would be sufficient to yield natural monopoly. But notice that if production costs exhibit decreasing returns, then natural monopoly is not necessarily implied. If decreasing returns in production increase costs more than network effects increase benefits, multiple incompatible networks (standards) can coexist.

If network effects are dominated by natural monopoly influences, and if tastes are relatively homogeneous, then only one standard will survive in the market. Given this all-or-nothing choice, an important economic question is whether markets will reliably chose the best standards among the available alternatives. In the usual public-utilities context for the natural monopoly problem, this issue does not arise, or does not arise with much force, because we typically have thought of the products of these natural monopolies as homogeneous. 'Plain Old Telephone Service' might have a quality dimension that could be regulated, but the product was not considered to differ in fundamental ways from one provider to another, at least not in economic modeling. The same could be said for electricity. Otherwise, it is generally just assumed that the natural monopolist who comes to dominate a market will be at least as efficient as any other producer.

The current literature on network effects departs from these assumptions, although specifics differ across the many models populating it. While the details differ from one model to the next, a common feature of a class of network models is that the expected result for decreasing-returns goods, that markets will chose the products that yield the greatest net benefits, is not assured. Consumers' expectations become extremely important. Consumers may choose a product that they would not choose except for an expectation that most other consumers will chose it. Within these models, markets cannot be relied upon to choose better networks or products.

In the alleged market failures, the forgone superior alternatives are held to offer gains even when switching costs are taken into account. Obviously, if the switching costs outweigh other benefits, it would not be efficient to switch. The claim in this literature is that even if it makes sense to switch to a better product, the switch might not take place.

While the presence of network effects may counter the usual expectation that markets choose products that provide the greatest surpluses, the mere existence of network effects and increasing returns is not sufficient to make possible the choice of an inferior technology. For that, some additional assumptions are needed. One

common assumption that can generate a prediction of inefficient network choice is that the network effect differs across the alternative networks (Arthur, 1989). In particular, it is sometimes assumed that the network offering the greatest surplus when network participation is large also offers the smallest surplus when participation is small. In other words, when the networks start out, network A provides greater benefits then network B. Users pick A, therefore, and this only makes A relatively more attractive, so that A comes to dominate. But B is actually a better network when there are large numbers of users.

This condition, however, is not likely to be satisfied in many real world instances, since synchronization effects are frequently uniform across networks. For example, if there is value in a cellular telephone network becoming larger, this should be equally true whether the network is digital or analog. Similarly, the network value of an additional user of a particular video recorder format is purported to be the benefits accrued by having more opportunities to exchange or rent videotapes. But this extra value should not depend on the particular format of video recorder chosen. If network effects are the same for all versions of a given product, it is very unlikely that the wrong format would be chosen if both are available at the same time.

Nevertheless, the theoretical possibility remains that markets with network effects might get locked-in to inferior products, even if the conditions under which this might happen appear restrictive. It becomes, then, an empirical issue whether this possibility happens with any frequency in the economy. There are some examples of lock-in that have been put forward by supporters of this theory.

The most famous empirical articulation of this view, which is extensively cited in the network effects literature, can be found in David (1985). David's paper puts forward the typewriter keyboard as an example of a market where network effects have caused consumers (typists) to become 'locked-in' to an inferior standard when there is a clear alternative that would make them better off, in this case the Dvorak keyboard. David notes several characteristics that might help lead to this result, chief among them being a 'system scale economy' that favors larger systems, his version of a network effect.

David provides what we now know to be a skewed version of the history of the QWERTY typewriter keyboard. David relies on reports from a World War II US Navy study that purportedly demonstrates that the costs of retraining typists on the Dvorak keyboard would be amortized within ten days of their subsequent full-time employment. In spite of such an advantage for Dvorak, it is not adopted. David concludes that coordination failures in the economy prevented firms and individuals from making the profitable investment of converting to Dvorak. David further tantalizes readers by suggesting that there are "many more QWERTY worlds lying out there." Similar stories are repeated or put forward in Arthur (1990), who adds the VHS video recorder format and the internal combustion engines as examples of QWERTY worlds gone awry.

Liebowitz and Margolis (1990, 1994) examine these stories in some detail and find that they were essentially myths. David, it turned out, had never actually examined the Navy study. A close inspection of the Navy study indicated various methodological errors that biased the results in favor of the Dvorak keyboard. The conditions under which the Navy study was conducted are quite murky, although it appears that they were conducted by the Navy's top expert in time and motion study, Lieutenant Commander August Dvorak, the inventor (and patent holder) of the alternative keyboard bearing his name. David also failed to discover a much better documented study by the General Services Administration that attracted a great deal of attention while it was being conducted in the 1950s That study found no advantage for Dvorak. Further, ergonomic studies of keyboards have found little difference between the two keyboards. Even studies performed by Dvorak disciples that purported to show some advantage for the Dvorak keyboard provided results which were much closer to that found in the General Services Administration study than those of the Navy study.

Examinations of other putative market malfunctions also fail to reveal evidence to support the lock-in claim. Beta, for example, came before VHS, and since the two companies sponsoring each format had a patent sharing agreement, the two formats differed mainly in the playing time they could provide. VHS, having a larger tape, could always, at any given speed or quality level, record or play for a longer time then could Beta. The original Beta did not even have the ability to record a movie since its taping time was limited to one hour. Suggestions that Beta might have had a better picture were found to lack any substantiation. Other examples that have been put forward have not served any better at supporting lock-in claims.

4. Features of modern network effects models

While a detailed accounting of each of the many network-effects models is beyond our scope here, there are certain assumptions that many network models hold in common. As we have noted, modern network-effect models often feature a particular set of outcomes; survival of only one network or standard, unreliability of market selection, and the entrenchment of incumbents. Not surprisingly these results follow from particular assumptions. In turn, these assumptions are simplifications that are often found in economic theory and that appear to be relatively unrestrictive. As applied in these network models, however, these assumptions are both responsible for the results and highly restrictive.

Natural monopoly is essential for the choice-of-networks problem that has been the focus of much of the network-effects literature. After all, if multiple networks can co-exist, then individual users can chose their most preferred

network, and the social discrete-choice problem does not arise. In fact, two of the assumptions common to most network-effects models are tantamount to natural monopoly. Network models have commonly assumed that marginal production cost is constant (or zero) and that network value functions that rise without limit. See, for example, Chou and Shy (1990, p 260); Church and Gandal (1993, p 246), Katz and Shapiro (1986, p 829), and Farrell and Saloner (1992, p 16). Matutes and Regibeau (1992) consider issues of duopoly and compatibility under a similar structure. Such assumptions impose an inexhaustible economy of large-scale operation. If, on the other hand, network size can reach a point where additional participation does not provide additional value to participants, then beyond that point, increases in scale are no longer advantageous and it is possible for multiple networks to compete, with each operating at an efficient output.

Without investigation, it seems unreasonable that in all or most new-technology industries the law of diminishing marginal product is somehow suspended. While the scale properties of a technology pertain to the simultaneous expansion of all inputs, it seems evident that resource limitations do ultimately constrain firm size. Economists have long supposed that limitations of management play a role in this, a relationship formalized in Radner (1992). The resource constraints faced by America Online in early 1997 is an example of this type of effect, even though it was only a short-run phenomenon at that time.

Another assumption common to most network models is that consumers are identical in their valuations of alternative networks. Once heterogeneous tastes are admitted to the model, it becomes feasible for competing networks to coexist with one another even though each exhibits natural monopoly characteristics. For example, if some computer owners much prefer Macintoshes and others much prefer the PC, they could coexist.

A further restriction in the modeling is the undifferentiated value received by consumers when another consumer joins a network, regardless of who the new consumer is. If economists, for example, much prefer to have other economists join their network as opposed to, say, sociologists, then a sociologist joining the network would have a smaller network effect than another economist. Such differential network impacts make it possible for economists to form a coalition that switches to a new standard even if the new standard failed to attract many sociologists. The possibility that groups of people could address efficient network choices becomes important when we consider the influence of large firms on the adoption of efficient standards. These corporate influences may be one reason that real-world examples of lock-in seem exceedingly hard to find.

All three assumptions lean in the direction of single-network equilibria for markets that exhibit network effects. In turn, such equilibria make possible, though certainly not necessary, the choice of a suboptimal network.

5. The empirical importance of network effects

5.1. Network effects in principle

Although many technologies have tended to evolve into single formats (e.g. home-use VCRs are almost all of the VHS variety) some portion of these may actually have evolved for reasons having little to do with either network effects or increasing returns. We should not be surprised to find that where there are differences in the performance of various standards, one may prevail over the others simply because it is better suited to the market.

First of all, the extent (and homogeneity) of network effects may be much more limited than is commonly assumed. For example, in the case of word-processors, it may be quite important for a small group of collaborators to use identical software so as to be perfectly compatible with each other. Similarly, compatibility may be important for employees within a firm. But compatibility with the rest of the world may be relatively unimportant; unimportant enough to be overwhelmed by differences in preferences, so that multiple networks could survive. Networks that serve niche markets well (such as word-processors specializing in mathematical notation) might not be significantly disadvantaged by network effects.

As an illustration, consider tax software and financial software. By far the dominant firm in North America is Intuit, with its Turbo-Tax products and Quicken financial software. This market seems to have tilted strongly toward a single producer. Yet network effects should be virtually nonexistent for these products. Consumers do not exchange this type of information with one another. A better explanation that is consistent with the product reviews in computer magazines is that these products are just better than the alternatives (Liebowitz and Margolis, 1999).

It is true that the past decades have evidenced a number of technologies that have experienced enormous declines in prices and tremendous growth in sales. Nevertheless, it is not clear that this is the result of increasing returns to (network) scale *per se*. Since bigger has been cheaper, it has often been assumed that bigger causes cheaper. But an available alternative explanation is that as technologies have advanced with time, the average cost curves (derived under *ceteris paribus* assumptions) have themselves been shifting down over time. If that is the case, the implied causality may be reversed: cheaper causes bigger. Consider for example the history of old technologies, such as refrigerators and automobiles (or a modern product such as the personal computer). These industries, currently thought to exhibit conventional decreasing returns in production (beyond a minimum efficient scale), experienced tremendous cost decreases, along with tremendous increases in utilization, early in their histories. Changes in technology may have been more important than changes in scale. (For a related discussion of increasing returns and monopolization, see Stigler, 1941, pages 68–76).

Liebowitz and Margolis (1999) present additional alternative explanations for single-product or single-firm domination of an industry. Production of software may be subject to instant scalability, which is to say that production can be substantially scaled up in a very short time because the hardware used to make reproductions is not specialized for a particular title. Thus if a product in a particular category of software is identified by consumers as the best in category, production can be scaled quickly to satisfy market demand. In studies of software markets, this explanation appeared to fit the behavior of market shares over time better than network effects or lock-in explanations. In addition to exhibiting very high market concentration, software markets exhibit extraordinary volatility of market shares and turnover of market leaders consistent with product reviews.

Though economists have long accepted the possibility of increasing returns, they have generally judged that, except in fairly rare instances, the economy operates in a range of decreasing returns. Some proponents of network externalities models predict that as newer technologies take over a larger share of the economy, the share of the economy described by increasing returns will increase. Brian Arthur has emphasized these points to a general audience: "[R]oughly speaking, diminishing returns hold sway in the traditional part of the economy – the processing industries. Increasing returns reign in the newer part – the knowledge-based industries... They call for different management techniques, strategies, and codes of government regulation. They call for different understandings" (Arthur, 1996: 101). Whether these markets require different management techniques, and what these techniques might be, is not yet clear.

5.2. Measuring network effects

Network effects have been the focus of only a very limited and somewhat tenuous empirical literature. In several instances, these examinations have foundered on how one might go about measuring a network effect. These conceptual issues are more difficult than one might think.

On the one hand, we have the rather obvious strength of network effects in literal networks such as telephones, facsimile machines, and automated teller machines. We suspect that no one would doubt the importance of network effects in the case of literal networks, which is the type of network that should be most relevant in the telecommunications arena.

On the other hand, we have the question of the importance of network effects in certain virtual networks, such as the network of software users. Because the case for virtual networks seems less certain, these are probably the more interesting topics for empirical examination.

We are aware of two attempts to measure the strength of network effects in virtual networks. Both are based upon software markets. Gandal (1994) and Brynjolfsson and Kemerer (1996) each use spreadsheet markets to test for network effects. Both run hedonic regressions with putative proxies for network

effects as independent variables. Gandal uses Lotus file compatibility and the ability to link to external databases to measure the strength of network effects. Brynjolfsson and Kemerer use Lotus menu structure and installed base as a measure of network effects.

In general, these variables do not measure network effects distinctly. The value attributed to Lotus file compatibility will be the result of two influences. The first is that consumers who are upgrading from older products will benefit from the ability of the new product to read their old files. This is not a network effect. The second influence is the network effect of compatibility with others. Similarly, the ability to link to external databases is a useful function, but it is not a network effect since it does not depend on the number of other 1-2-3 users. The Lotus menu structure variable measures the importance of remaining compatible with one's old learned habits, but this too is not a network effect[4]. The installed base variable suffers from a similar problem, since if there were no new buyers (but only upgrade buyers) it might merely pick up the value of being compatible with old files. These variables might have measured the effects that they were intended to measure if the sample of users were limited to first time buyers, but they were not.

The empirical importance of virtual network effects in any specific market, therefore, is still an open question. The empirical importance of network effects in literal networks may seem so obvious that empirical testing may have seemed redundant. Nevertheless, there have been several recent attempts to measure or test for network effects in the case of the facsimile market and automated teller machines.

Saloner and Shepard (1995) attempt to measure the relative importance of network effects and scale economies in the adoption of ATMs (automated teller machines). The idea is that consumers receive additional value from having access to a large network of ATMs rather than a small network, certainly a reasonable assumption. Their time period was the decade of the 1970s, however, a period prior to interconnected ATM networks of multiple banks that are common now. They find that the probability of a bank installing at least one ATM is positively related to the number of branches, evidence that they believe confirms the importance of network effects.

Although their results are expected and entirely plausible, their analysis can not be considered to have confirmed the hypothesis due to some very basic problems with their estimation. For one thing, random adoption of ATMs by bank branches would also create a positive measured relationship, although Saloner and Shepard test for this randomness effect and find its impact limited. Second, Saloner and Shepard find that the relationship does not hold for the specific states, within the United States, that do not restrict the number of branches that banks

[4] Except for the unlikely case where it is important to consumers that other random individuals be able to borrow their computers and easily use their spreadsheet program.

can offer. Since this sample would minimize the likelihood of conflating economic and regulatory impacts, this failure throws the other results into some doubt. Saloner and Shepard explain this puzzling result by demonstrating that multicollinearity plagued the data.

Finally, and most importantly, the measurement of network size that they used is the number of bank locations, not the number of free standing ATMs. This is because ATMs, according to Saloner and Shepard, were virtually always contained in branches during the period they analyze. This is a serious problem, for several reasons, foremost among them being that it is not clear under these conditions that the tests performed by Saloner and Shepard are capable of measuring literal network effects.

First, multiple-branch banks already exhibit virtual network effects, without the advent of ATM machines, assuming that the number of branches (or ATMs for that matter) are related to the number of users. Consumers who choose banks with many branches are likely to value the network effects far more than consumers who choose banks with few or but a single branch. The co-existence of single and multiple branch banks practically guarantees that there are wide variations in the importance of network effects to consumers, and that there was probably an important self-selection bias before the introduction of ATMs. Since consumers who use single branch banks presumably have little value for other bank locations, the value they get from an ATM should have nothing to do with the number of other ATM locations. The value they get from an ATM depends only on any other advantages it brings, such as nighttime transactions or a shorter wait. Given that multiple branch banks already have customers who value the ability to go to multiple branches, the value they get from ATMs is also only the advantage ATMs bring to an already networked bank, such as the night-time transactions or shorter waits. The adoption of ATMs will depend on the relative value these two groups place on the ATM advantage, not on any difference in network effects.

A more robust examination of network effects is found in Majumdar and Venkataraman (1998). They examine, among other things, the impact of network density (a proxy for network effects) on the adoption of new switches in the US telecommunications industry. They find that density and heterogeneity in user tastes enhances the adoption speed of these technologies, technologies that allowed new forms of interactions among users. Their results, although not very strong, support the view that the more other users there are that can use these technologies, the more valuable these technologies will prove to users. Of course, in these literal networks it is to be expected that network effects should be important and the difficulty in finding strong support only illustrates the empirical problems involved.

Economides and Himmelberg (1995) attempt to measure the network effects in the introduction of facsimile machines. We do not believe anyone would doubt that network effects are important for fax machines. One interesting question,

however, is whether the explosion in facsimile machine sales that occurred in the mid-1980s was due to rapidly declining price, or instead to the network effects brought about by the expanded base of users. While Economides and Himmelfarb conclude that network effects are the key element, the complex structure of their estimated coefficients, and the fact that they only had 14 observations, makes it difficult to have much confidence in their results.

6. Policy implications

Obviously, telecommunication policies, such as the optimal pricing for state run telephone companies, should have been and were influenced by the earlier thinking about network effects. The current policy implications of this literature are now focused, however, on the modern version of network effects, in particular the ideas of lock-in and path dependence.

For example, the AOL/Time-Warner merger plans were heavily scrutinized by regulatory agencies on both sides of the Atlantic. The European Commission, which did ultimately approve the merger, raised a serious concern that AOL/Time-Warner might become an entrenched monopolist in the delivery of music and other online content, as yet a barely nascent market. A report by the European authorities referred specifically to network effects described as: "More subscribers bring more content, and vice versa ... the more content AOL acquires and the bigger its community of users, the less the reasons for a user to abandon AOL's walled garden, and the greater the reasons for potential Internet users to join AOL"[5]. In response to this concern, a planned merger between Time-Warner and EMI was called off. Yet, just because music is sold *over a network* does not mean that there are network effects involved. Indeed, there are few network effects that can be imagined in the sale of music online.

American antitrust authorities also were concerned about the merger and network effects. Although their major concern had to do with AOL/Time-Warner providing cable access to ISPs, American regulators were also disturbed about the dominant market share of the AOL messaging service. Because these messaging services have strong literal network effects, regulators feared that it might remain an entrenched monopolist as might be predicted from a naive reading of the lock-in literature.

The theory of network externality played a crucial role in the American government's antitrust prosecution and the district court's findings in the Microsoft case[6]. The judge referred repeatedly to Microsoft's 'application barrier to entry' which is just another term for the indirect virtual network effect that might be

[5] The Wall Street Journal "Time Warner, AOL are criticized by European Antitrust Enforcers," John R. Wilke, August 31, 2000.

[6] United States v. Microsoft Corp., 1995, 159 F.R.D. 318, 333-38, D.D.C., 1995, Sporkin, J., rev'd, 56 F.3d, 1448, D.C. Cir.

associated with operating systems. The claim is that Windows, having 70,000 applications, has enormous indirect network effects that its competitors such as the Macintosh, with only 12,000 applications, cannot match. The claim was that these network effects were the key to Microsoft's monopoly position and that Microsoft broke the law in trying to protect Windows' network effects. The power of the network effects story is such that the district court has ruled that Microsoft be broken in half, although that decision is currently under appeal. Additionally, part of the government's claim, which would be virtually impossible to disprove, was that Microsoft decreased the level of innovation that might have occurred, a minor variation of the wrong-standard-winning argument.

Clearly the potential to misuse such as-yet unsubstantiated theories in antitrust actions, particularly for competitors unable to win in the marketplace, is very great, not unlike that of various theories of predation. With so little empirical support for these lock-in theories, it appears to us at best premature and at worst simply wrong to use them as the basis for antitrust decisions.

It is also possible that network effects may cast in a new light the role of copyright and patent laws. First, networks are likely to be too small if network effects are not internalized. Intellectual property laws are one means by which such network effects can be internalized, since ownership is usually an ideal method of internalization.

Further, where one standard is owned and another is not, we can have less confidence that an unowned but superior standard will be able to prevail against an owned standard, since the owner of a standard can appropriate the benefits of internalizing any network effects. Although we do not have any evidence that inferior standards have prevailed against superior standards (in free markets), this may be in large part because most standards are supported by companies that have some form of ownership such as patent, copyright, or business positioning. The greatest chance for some form of remediable market failure would be if an unowned standard with dispersed adherents were to engage in competition with a standard that had concentrated ownership. Additional research in this area is needed, however, before any firm conclusions can be drawn.

7. Conclusions

Network effects and other consumer-side economies and diseconomies have long been a subject of interest in telecommunications economics and economics in general. In fact, one of the central features of telecommunications policies, and debates over those policies, is universality of service, a concern that arises in part from network effects. The arrival of new telecommunications technologies, and the seemingly profound influence of these new technologies on commerce and consumption, has focussed attention on the existence and consequences of these effects.

Network effects raise two sorts of concerns about the outcomes of decentralized decision marking of laissez faire markets. The first relates to the sizes of networks, the second to the choice of a network among competing alternatives. The first of these potential market failures can arise if network participants do not capture the benefits that their participation confers on others. The second arises because scale economies may result in the survival of only one network among the conceivable alternatives, which in turn may prevent decentralized individual choices from being properly counted in the market. Each of these failures has been characterized as a kind of externality. So strong is the predisposition to find market failure that the network interaction that the first writings in the recent literature on networks defined these effects as network externalities. Since then the more neutral term 'network effects' has become common, with network externality now recognized as being a special case.

Network effects will be internalized in many empirically important circumstances. Where networks are owned, owners have an incentive, and in important cases the means, to internalize these effects. Such internalization can address both the size of network and the choice among competing networks.

The empirical literature on network effects is new and so far quite limited. Efforts to evaluate the empirical importance of network effects have been hampered by limited data, inadequate definition of the phenomenon, and other methodological problems. Nevertheless, measurement of these effects is likely to be worthwhile, given the widespread claims of the importance of these effects and the important role that such claims are beginning to play in public policy.

References

Arthur, W. B., 1989, Competing technologies, increasing returns, and lock-in by historical events, Economic Journal, 99, 116–131.

Arthur, W. B., 1990, Positive feedbacks in the economy, Scientific American, 262, 92–99.

Arthur, W. B., 1996, Increasing returns and the new world of business, Harvard Business Review, 72, 100–109.

Artle, R., and C. Averous, 1973, The telephone system as a public good: Static and dynamic aspects, Bell Journal of Economics, 4, 89–100.

Besen, S. M. and J. Farrel, 1994, Choosing how to compete: Strategies and tactics in standardization, Journal of Economic Perspectives, 8, 117–31.

Brynjolfsson, E. and C. F. Kemerer, 1996, Network externalities in microcomputer software: An econometric analysis of the spreadsheet market, Management Science, 42, 1627–1647.

Chou, D. and O. Shy, 1990, Network effects without network externalities, International Journal of Industrial Organization, 8, 259–270.

Church, J and N. Gandal, 1993, Complementary network externalities and technological adoption, International Journal of Industrial Organization, 11, 239–260.

David, P. A., 1985, Clio and the economics of QWERTY, American Economic Review, 75, 332–337.

Economides, N., 1996, The economics of networks, International Journal of Industrial Organization, 14, 673–700.
Economides, N. and C. Himmelberg, 1995, Critical mass and network evolution, in G. Brock, ed., Toward a Competitive Telecommunications Industry: Selected Papers from the 1994 Telecommunications Policy Research Conference, Mahwah, NJ: Lawrence Earlbaum Associates.
Ellis H. S. and W. Fellner, 1943, External economies and diseconomies, American Economic Review, 33, 493–511.
Farrell J. and G. Saloner, 1985, Standardization, compatibility, and innovation RAND Journal of Economics, 16, 70–83.
Farrell J. and G. Saloner, 1992, Converters, compatibility, and control of interfaces, Journal of Industrial Economics, 40, 9–36.
Gandal, N., 1994, Hedonic price indexes for spreadsheets and an empirical test of the network externalities hypothesis, RAND Journal of Economics, 25, 160–170.
Kahn, A. E., 1988, The economics of regulation: Principles and institutions, Cambridge, MA:MIT Press.
Katz M. L. and C. Shapiro, 1985, Network externalities, competition, and compatibility, American Economic Review, 75, 424–440.
Katz, M. L. and C. Shapiro, 1986, Technology adoption in the presence of network externalities, Journal of Political Economy, 94, 822–841.
Katz, M. L. and C. Shapiro, 1994, Systems competition and network effects, Journal of Economic Perspectives, 8, 93–115.
Knight, F. H., 1924, Some fallacies in the interpretation of social cost, Quarterly Journal of Economics, 38, 582–606.
Leibenstein, H., 1950, Bandwagon, snob, and Veblen effects in the theory of consumer's demand, Quarterly Journal of Economics, 64, 183–207.
Liebowitz, S. J. and S. E. Margolis, 1990, The fable of the keys, Journal of Law and Economics, 33, 1–26.
Liebowitz, S. J. and S. E. Margolis, 1994, Network externality: an uncommon tragedy, The Journal of Economic Perspectives, 8, 133–150.
Liebowitz, S. J. and S. E. Margolis, 1995a, Path dependence, lock-in and history, Journal of Law, Economics and Organization, 11, 205–226.
Liebowitz, S. J. and S. E. Margolis, 1995b, Are network externalities a new source of market failure? Research In Law And Economics, 17, 1–22.
Liebowitz, S. J. and S. E. Margolis, 1996, Market processes and the selection of standards, Harvard Journal of Law and Technology, 9, 283–318.
Liebowitz, S. J. and S. E. Margolis, 1999, Winners, losers, and Microsoft: competition and antitrust in high technology, Oakland, CA: Independent Institute.
Littlechild, S. C., 1975, Two-part tariffs and consumption externalities, Bell Journal of Economics, 6, 661–670.
Majumdar, S. K. and S. Venkataraman, Network effects and the adoption of new technology: Evidence from the U.S. telecommunications industry, Strategic Management Journal, 19, 1045–1062.
McKie-Mason, J. W. and H. R. Varian, 1995, Pricing the Internet, in B. Kahin and J. Keller, eds, Public Access to the Internet, Cambridge, MA: MIT Press.
Matutes, C. and P. Regibeau, 1992, Compatibility and bundling of complementary goods in a duopoly, Journal of Industrial Economics, 40, 37–54.
The Online Wall Street Journal, 2000, Time Warner, AOL are criticized by European antitrust enforcers, John R. Wilke, August 31.
Radner, R., 1992, Hierarchy: The economics of managing, Journal of Economic Literature, 30, 1382–1415.
Rae, J., 1905, The sociological theory of capital, London: Macmillan Company.

Rohlfs, J., 1974, A theory of interdependent demand for a communications service, Bell Journal of Economics, 5, 16–37.
Saloner, G., and A. Shepard, 1995, Adoption of technologies with network effects: An empirical-examination of the adoption of automated teller machines, RAND Journal of Economics, 26, 479–501.
Squire, L., 1973, Some aspects of optimal pricing for telecommunications, Bell Journal of Economics, 4, 515–525.
Stigler, G. J., 1941, Production and distribution theories, New York: Macmillan Company.
Young, A. A., 1913, Pigou's wealth and welfare, Quarterly Journal of Economics, 27, 672–86.

Chapter 4

CUSTOMER DEMAND ANALYSIS

LESTER D. TAYLOR

University of Arizona

Contents

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

While the upheavals that have occurred in world telecommunications markets have generated many benefits, few of these have fallen on telecommunications demand analysts. Twenty years ago, the boundaries of the telecommunications industry were stable and well-defined, as were the services provided and telephone companies were either state-owned or regulated monopolies, which made for a readily available body of data on a consistent and comprehensive basis. Wireless telephony was in its infancy, and the Internet was scarcely a rumor. Now, all is different. Industry boundaries are in a constant state of flux, telecommunications markets are increasingly competitive, wireless and the Internet have changed the way that many communicate, cable access may be poised to make a major presence, and company-based data have become increasingly fragmented and proprietary. As a consequence, telecommunications demand analysts must now contend with a rapidly changing and expanding mix of services, take into account emergent substitutes and complements, deal with firm demand functions as opposed to industry demand functions, and collect and organize primary data. As one who has been extensively involved in building models of telecommunications demand since the mid-1970s, I can state unequivocally that telecommunications demand analysis was much easier 20 years ago than it is today.

2. Telecommunications demand analysis in the 1970s and 1980s

As perspective on current tasks and challenges, a brief overview of the roles that telecommunications demand analysis has played in the past (with a focus on the U.S. and Canada) may be useful. The first substantive use of telecommunications price elasticities appeared in the early to mid-1970s, largely as a consequence of inflation and the slowing of technological change in long-distance transmission. During the 1950s and 1960s technological change had materially reduced the cost of inter-city transmission, which resulted in a steady reduction in long-distance rates. Demand analysis played virtually no role in the setting of rates. While rates were decreased, they were not reduced as much as the decrease in costs because of a desire on the part of regulators to maintain artificially low local rates.

The 1970s were different. The slowing of technological change in reducing the cost of long-distance transmission coupled with inflation caused the pressure on long-distance rates to be upward, rather than downward. Despite a strong desire on the part of regulators to keep local rates low, pressures were mounting to increase local rates as well. In this situation, the importance of price elasticities became apparent. With decreasing rates, telephone companies have little incentive to take price elasticities into account, since (with inelastic demand) revenues will be higher if price elasticity is ignored than if it is taken into account. The incentives clearly change, however, with upward adjusting rates, since revenues will

be lower if price elasticity is ignored than if it is not. As a consequence, the long-distance companies (AT&T in the US and the Trans Canada Telephone System in Canada) began to use econometrically estimated price elasticities in their filings for rate increases before the their respective regulatory commissions.

During the 1970s the focus was almost entirely on the estimation of toll elasticities, for this was where the action was in rate hearings. Access elasticities received little attention, as local rates were set residually to make up the difference between an agreed-upon revenue requirement and the revenues that would be projected to be forthcoming in the toll and other 'actively priced' markets. By the late 1970s sophisticated toll demand models for the interstate US market existed at AT&T and the FCC and for the Canadian inter-provincial market at Bell Canada, and intrastate models had been used in rate hearings in more than three-quarters of the US states[1]. In great part because of the role played by demand modelers at AT&T, a generic toll demand model emerged in the, then, Bell System. In that model a measure of toll calling activity, usually calls or minutes, was related to income, price, a measure of market size, typically the number of telephones, and 'habit' which was usually measured by the value of the lagged dependent variable.

By the late 1970s things were again in a state of flux in telecommunications markets. Competition had emerged in the U.S. in the intercity toll market and the historic toll-to-local subsidies that had sustained artificially low local rates were being eroded as a consequence of competition driving toll rates towards cost. The result was substantially increased upward pressure on local rates, together with an extensive exploration of the possibilities of replacing flat-rate local service with measured service. Amid widespread concern that increased local rates and/or the substitution of measured for flat-rate local service would cause large numbers of households to give up telephone service – and therefore jeopardize universal service – attention in the early 1980s shifted to the estimation of access price elasticities.

The estimation of access demand models required the use of different modeling strategies from that of estimating toll models. Whereas the 1970s toll models were usually estimated by conventional regression methods using market data, access demand models were usually specified in a probit or logit discrete choice framework and estimated from data referring to individual households. Moreover, interest was not only on the overall magnitude of access price elasticities, but how access elasticities differed between and among low- and high-income households, black, Hispanic and white households, and single-parent households headed by females[2].

The divestiture of AT&T that went into effect on January 1, 1984 was a significant event not only for the restructuring of the telephone industry in the US,

[1] For a summary of these models, see Taylor (1994, Appendix 2).

[2] A generic form of the access demand models that were estimated in the 1980s for the U.S. and Canada will be discussed below.

but for telecommunications demand analysis as well. Historically, data relating to the telephone industry had been conveniently collected and made freely available to the public by AT&T. With the divestiture, however, this was no longer the case, as AT&T (which still formed the overwhelming bulk of the intercity toll market) quickly came to view its data as proprietary. Demand analysts accordingly now had to expend a great deal more effort in collecting and organizing data sets.

The divestiture eliminated AT&T as a primary source of the data, and it underscored the fact that the Bell System was no longer co-terminous with the telephone industry. No single company, even if it were so willing, could be a provider of data for the entire industry. In short, by the mid-1980s, not only had data in the telephone become proprietary, but it had become severely fragmented as well. Moreover, with competition and fragmentation, firm elasticities became distinct from industry elasticities, and care had to be used in interpreting elasticities that had been estimated with the data of a single company.

An additional consequence of the divestiture was that it restructured the intercity toll market into three markets in place of the two that had previously existed. Pre-divestiture, there was an intrastate market and an interstate market. State public utility commissions regulated rates in the intrastate markets, while interstate rates were regulated by the FCC. With divestiture, the interstate market remained intact, but the intrastate market was split into two parts, an intraLATA market that was restricted initially to local exchange companies and an interLATA/intrastate market that was open to any long-distance company, but was closed to the local exchange companies. The FCC continued to regulate interLATA/interstate rates, while intraLATA and interLATA/intrastate rates remained under jurisdiction of the individual states.

With competition in the interLATA markets, the need to distinguish between firm and industry elasticities was immediately evident. Initially, this was not a problem in the intraLATA market, but in time competition emerged in that market as well, both as a result of policy, and benignly through the use of 1-0-XXX (or now 10-10-XXX) dialing[3]. As this occurred, the same problems of analysis emerged as in the interLATA markets. In fact, the problems were even more insidious with intraLATA data because analysts, using data obtained from the local exchange companies, were often not even aware that the data did not pertain to the entire intraLATA market[4].

[3] The XXX refers to the unique three-digit code that was assigned each of the long-distance carriers. Thus, 1-0-XXX for AT&T was 1-0-ATT (or 1-0-288). Dialing this number would connect the dialer directly to the AT&T toll switch, through which an intralata call could then be placed as though it were an interlata call.

[4] Objectively, there is no way of knowing just how much the intralata price elasticities estimated using data from the billing records of local exchange companies may have been affected as estimates of the industry elasticity for the intralata market. Ways of dealing with this problem using data obtained from customers, rather than telephone companies, will be discussed below.

Another demand-modeling problem that emerged after the AT&T divestiture was a need to deal with the so-called custom- and class-calling features offered by the local telephone companies. In practice, services such as call waiting, call forwarding, three-way calling and speed dialing came to be offered individually or in packages of any two, any three, or all four. The bundled (or packaged) price was of course less than the cost had the consumer purchased the items in the bundle individually. In the early 1990s several approaches to the modeling of the demand for these bundled services were pursued, including regular logit, conditional (or nested) logit, and multinomial probit models.

3. The general nature of telecommunications demand

The feature that most distinguishes telecommunications demand from the demand for most goods and services is the fact that telecommunications services are not consumed in isolation. A network is involved. This gives rise not only to certain interdependencies and externalities that affect how one models consumption, but also creates a clear-cut distinction between access and usage. There must be access to the network before it can be used. How this distinction has conventionally been modeled in the literature will be dealt with in Section 4.

The interdependency that has received the most attention in the literature is what is usually referred to as the *network* (or *subscriber*) externality. This externality arises when a new subscriber joins the network. Connection of a new subscriber confers a benefit on existing subscribers because the number of ports that can now be reached is increased. As this benefit is shared in common by all existing subscribers, the network externality accordingly has the dimension of a public good. The importance of this externality lies in its potential to generate endogenous growth. Because a larger network is more valuable to belong to than a smaller network, an increase in network size may increase the willingness-to-pay of existing extra-marginal subscribers to the point where they become actual subscribers. This in turn may induce a succeeding set of extra-marginal subscribers to subscribe, and so on and so forth[5]. However, the possibility of this phenomenon being of any present importance in mature telephone systems, as in North America and Western Europe, is obviously remote[6].

[5] The treatment of consumption externalities in telecommunications literature long predates the emergence in the 1990s of a voluminous literature on the economics of networks. For a review of the latter literature, see the contribution to this *Handbook* of Liebowitz and Margolis. The seminal papers in telecommunications are Artle and Averous (1973), Rohlfs (1974), Von Rabenau and Stahl (1974) and Littlechild (1975).

[6] In the U. S. and Canada, the time (if ever) for network externalities to have been a factor in fueling endogenous growth in the telephone system would have been in the first decades of the twentieth century. On the other hand, they may have been a factor in France in the 1970s, and are almost certainly of prime importance in some present-day undeveloped countries. The Internet, of course, is another matter, and there seems little question but that network externalities have been (and continue to be) of great importance in its phenomenal growth.

Additional externalities are associated with usage. A completed telephone call requires the participation of a second party, and the utility of this party is thereby affected. The gratuitous effect which falls on the second party represents what is usually referred to the *call* (or *usage*) externality. Unlike the network externality, the call externality has never received much attention, most probably because it has never been seen as creating a problem for pricing. Assuming that the call externality is positive, recipients of calls (since most calls are paid for by the originator) receive unpaid-for benefits. The existing view in the literature pretty much has been that, since most calling is bi-directional, the call externality is internalized over time as today's recipients of calls become tomorrow's originators. This balances out obligations *vis-a-vis* one another. For the system as a whole, the benefits from incoming calls can be seen as increasing willingness-to-pay to have a telephone, which means that a telephone system will be larger when call externalities are present than what it would be in their absence.

The foregoing is the conventional view of the call externality and is uncontroversial[7]. But it is also not very useful empirically, for its only implication is that a telephone system will be larger with call externalities than without. Nothing is said or implied about usage, which seems strange since the source of the externality is in fact usage. The problem is that an important aspect of the interdependency between callers is being overlooked. To see what is involved, let consider two individuals, A and B, who regularly communicate with one another by telephone and for each of whom the call externality is positive, in the sense that each of them looks forward to receiving a call from the other. Two circumstances can be identified. One in which the number of calls between A and B is determined 'exogenously' as a function (say) of their respective incomes and the price of calls. A second circumstance in which the number of calls between the two depends not only on income and price, but also on the number of calls that A makes to B and the number of calls that B makes to A. Calling is clearly endogenous in the second circumstance.

In the first circumstance, one can easily imagine the appearance of an implicit contract in which A calls B half of the time and B calls A for the other half[8]. This is the situation envisioned in the conventional interpretation of the usage externality. It simply involves the internalization of the externality through an implicit contract. Income and price determine the total number of calls. The only effect of the externality is to help determine how the cost of the calls is divided.

The second circumstance is clearly much more complex, because not only can there be an implicit contract defining 'turn,' but it is a situation in which a call can create the need for further calls. The call externality in this case not only helps to determine who pays for calls, but also leads to an actual stimulation of calls. An example of this would be where A's call to B increases B's utility so much that she

[7] See Squire (1973) and Littlechild (1975).
[8] See Larson and Lehman (1986).

calls A back to tell him so, or else calls C to tell her that A had called. Common as this event is in real life, a more frequent phenomenon (especially among business users) is probably one in which further calling is not rooted in the call externality as conventionally defined. This is the situation in which an exchange of information *creates the need* for further exchanges of information. Several authors collaborating on a paper provides an obvious example. For lack of a better term, I have referred to this phenomenon as the *dynamics of information exchange*[9].

The dynamics of information exchange may not only involve B returning a call to A as a result of an initial call from A, but may create a need for B to call C to corroborate something that A had told B, which in turn may lead to C calling D, and so on. Essentially the same dynamics can operate whenever, by whatever means, a piece of information is injected into a group that forms a community of interest. An obvious example is the death of one of the members of the group. A learns that B has died, calls C, and the grapevine kicks into operation[10]. A generic model that takes into account both this and the more conventional form of the call externality will be discussed in Section 4.

A further complication in modeling telecommunications demand arises from the fact that benefits arise not only from completed communications, but also from those that may not be made. Subscribing to the telephone network can be viewed as the purchase of options to make and receive calls. Some of the options will be exercised with certainty, while others will not be exercised at all. This is because many calls will only be made contingent upon the realization of random states of nature, and thus not known at the times that access is purchased. Emergency calls, such as for fire, police, or ambulance, are obvious cases in point, but compelling urgency is not the only determinant. Many calls arise from nothing more than mood or whimsy. *Option demand* is thus an important characteristic of telecommunications demand[11].

Reference to this point has for the most part been with respect to households and residential demand. Business demand is another matter, and unfortunately much more complicated. The complications include:

The many forms of business access to the telecommunications network. Fact one would seem to be that no business in this day and age can be without telephone

[9] Taylor (1994, Chapter 9).

[10] The externalities associated with the dynamics of information exchange may be of even greater importance with the Internet, whereby a visit to one website may create the need for visits to a whole string of other websites.

[11] The concept of option demand is well-known in the literature on conservation and exhaustible natural resources see Weisbrod (1964) and Kahn (1966), but apart from brief mention in Squire (1973), it is absent from the pre-1980 literature on telecommunications demand. In recent years, its relevance has come to the fore in analyzing the pricing of default capacity in the context of carrier-of-last-resort obligations and in the demand for optional extended-area service. See Weisman (1988) and Kridel, Lehman, and Weisman (1991). Access to the Internet over telephone lines provides yet another currently important impeller of option demand.

service. Accordingly, to approach business in an access/no access framework as with residential demand is simply not relevant[12]. What is relevant is the type of access that a business demands, together with the number of lines.

Internal vs. external communications needs. Unlike a household, much of business telecommunications demand is driven by internal needs. These needs vary depending upon the size and type of business, and increase rapidly with the number of locations. For large national and international corporations, the demand for internal communications probably outweighs the external demand. Large private networks are a reflection of this. The presence of multiple locations, perhaps more than any other factor, is what makes modeling business demand so difficult, for one must focus on communications between and among locations as well as between the firm and the outside world.

The importance of telecommunications in marketing. Another factor complicating analysis of business telecommunications demand is the heavy use of telecommunications in marketing. For many businesses, the telephone is a primary instrument of selling, whether it be in terms of direct communications with customers or maintaining contact with a sales force in the field.

Customer-premises equipment. During pre-competition in the telecommunications industry, modeling demand for customer-premise equipment was not difficult. Equipment was basic, it had to be rented, and its price was for the most part bundled into the monthly service charges. Access and equipment were essentially one and the same, and to explain the demand for one was pretty much to explain the other. Technological change and competition changed this. Access and equipment were unbundled, and customers were allowed to purchase or rent equipment from whomever they chose. As a consequence of this change, customer-premise equipment has become one of the most problem-plagued areas to model. Equipment is no longer slave to the choice of access, but is a decision variable of stature commensurate with access and usage.

The opportunity cost of time. An obvious consequence of economic growth is that time becomes more valuable. Businesses (as well as individuals acting in their roles of consumers) seek ways to make more efficient use of time, and often turn to increased use of telecommunications to do so. While this may seem paradoxical, since time spent on the telephone is real time generally unavailable for any other use, the point is obvious. Consider the savings that may arise from replacing sales people in the field with advertising, facsimile, e-mail and 800 services, the reduction of travel through teleconferencing, the use of the now near-universal ability to communicate with anyone from anywhere, and telecommuting. In most situations the relevant cost of increased telecommunications usage is

[12] This is not to argue that all business necessarily have telephones, but just as there are high-income households without telephone service, there are undoubtedly successful business for which the same is true. The numbers cannot be large, however, and the reasons are almost certainly non-economic.

not the out-of-pocket cost of the increased usage, but the cost of the labor and other resources that are replaced. Identifying and measuring these costs can, of course, be tricky.

Emergence of the Internet. Whatever the underlying structure of business telecommunications demand might be, it has almost certainly been juggled in recent years by the emergence of the Internet. While it is obvious that many telephone communications are lost to e-mail, it is not clear that ultimate impact of e-mail on telephone traffic is necessarily negative. Similar opaqueness applies to e-commerce and e-business. In my view, the best strategy for dealing with the impact of the Internet in modeling business telecommunications demand will be to encompass the Internet within an expanded definition of telecommunications.

4. Theoretical considerations

Let me turn to a brief review of the theoretical considerations that have guided the modeling of telecommunications demand since the mid-1970s. As noted in the preceding section, the point of departure is the distinction between access and usage.

4.1. A generic model of access demand

For notation, let a denote access, q a measure of calling activity, and CS a measure of the net benefits from usage. For convenience, assume that usage is measured by the number of calls at a price of p per call. The conventional approach to modeling the demand for access in the literature is in a discrete choice framework. Within that framework a is postulated to be a dichotomous (i.e., zero-one) random variable whose value depends upon the difference between the net benefits from usage of the telephone network (conditional on access) and the price of access, which will be denoted by the Greek letter π. The probability that a household will demand access can accordingly be written as

$$\begin{aligned} P(access) &= P(a = 1) \\ &= P(CS > \pi). \end{aligned} \tag{1}$$

Net benefits, CS, in turn, are usually measured in the literature by the consumer's surplus associated with the consumption of q, i.e., by

$$CS = \int_p^\infty q(z)\mathrm{d}z. \tag{2}$$

Perl (1983) was one of the first researchers to apply this framework empirically, doing so by assuming a demand function of the form[13]:

$$q = A\mathrm{e}^{-\alpha p} y^{\beta} e^{u} \tag{3}$$

where y denotes income and u is a random error term with distribution $g(u)$. Consumer's surplus, CS, will then be given by

$$\begin{aligned} CS &= \int_{p}^{\infty} A\mathrm{e}^{-\alpha z} y^{\beta} \mathrm{e}^{u} \mathrm{d}z \\ &= \frac{A\mathrm{e}^{-\alpha p} y^{\beta} \mathrm{e}^{u}}{\alpha}. \end{aligned} \tag{4}$$

With net benefits from usage and the price of access expressed in logarithms, the condition for demanding access to the telephone network accordingly becomes

$$\begin{aligned} P(\ln CS > \ln \pi) &= P(a - \alpha p + \beta \ln y + u > \ln \pi) \\ &= P(u > \ln \pi - a + \alpha p - \beta \ln y), \end{aligned} \tag{5}$$

where $a = \ln A/\alpha$.

The final step is to specify a probability law for consumer's surplus, which in view of the last line in Equation (5), can be reduced to the specification for the distribution of u in the demand function for usage. An assumption that u is distributed normally leads to a standard probit model, while an assumption that u is logistic leads to a logit model. Empirical studies exemplifying both approaches will be reviewed in Section 5[14].

4.2. *Models of toll demand*

As was noted earlier, by the late 1970s a generic model for toll demand had been developed and used extensively in the (then) Bell System, which included price,

[13] At the time that Perl's study was commissioned by the (then) Central Services Organization of (the about to become independent) Regional Bell Operating Companies, there was a lot of concern about the possible negative effects on telephone penetration of measured local service. This function, which has subsequently become known as the 'Perl' demand function, was motivated by the fact that it can accommodate a zero, as well as a non-zero, price for usage. There was also a lot of concern regarding the effects that higher local-service charges might have on subscription rates of non-white households and households headed by females. These effects were allowed for through a bevy of socio-demographic dummy variables.

[14] Most empirical studies of access demand that employ a consumer-surplus framework focus on local usage, and accordingly ignore the net benefits arising from toll usage. Hausman, Tardiff, and Bellinfonte (1993) and Erikson, Kaserman, and Mayo (1995) are exceptions.

income, market size, and habit as predictors. Market size was usually measured by the number of telephone access lines, while habit was represented by the lagged value of the dependent variable[15]. In the early 1980s a more sophisticated version of this model was developed by Bell Canada for use in hearings before the Canadian Radio-Television and Telecommunications Commission (CRTC). The ideas underlying the Bell Canada model are as follows.

Assume that there are two exchanges, A and B, that are not part of the same local-calling area, so that calls between the exchanges are toll calls. Let the number of telephones in the two exchanges be T and R, respectively. The total possible collections between the two exchanges will therefore be $T \cdot R$. Let M denote the number of calls that are 'sent paid' from A to B during some period of time (a quarter, say), and let θ denote the proportion of potential connections $(T \cdot R)$that are realized, so that

$$M = \theta\ (T \cdot R). \tag{6}$$

Equation (6) might describe the relationship that would be observed in a stationary world where income, price, and all other factors that affect toll calling are constant. However, these other factors do not remain constant, and we can allow for them through the value of θ. In particular, let us suppose that income (Y) and the price (P) of a toll call from A to B affect θ according to

$$\theta = aY^{\beta}P^{\gamma}, \tag{7}$$

where a, β, and γ are constants. The relationship in expression (6) accordingly becomes

$$M = aY^{\beta}P^{\gamma}(T \cdot R). \tag{8}$$

Finally, let suppose that M is affected by a change in potential toll connections that may be either more or less than proportional, in which case the model becomes

$$M = aY^{\beta}P^{\gamma}(T \cdot R)^{\lambda} \tag{9}$$

where λ is a constant, presumably of the order of 1. Taking logarithms of both sides of Equation (9), we obtain

[15] In the late 1970s various versions of this generic model had been estimated by local Bell Operating Companies for use in rate hearings in 34 states. A tabulation of the models can be found in Appendix 2 of Taylor (1994).

$$\ln M = \alpha + \beta \ln Y + \gamma \ln P + \lambda \ln(T \cdot R), \tag{10}$$

where $\alpha = \ln a$. With addition of a random error term, this equation represents the model that was used by Bell Canada in a variety of hearings before the CRTC in the 1980s and early 1990s

4.3. *Point-to-point toll demand*

Competition in the interLATA toll market in the U.S. in the 1980s quite naturally stirred interest in disaggregation and the development of toll demand models that were route-specific[16]. Models were developed that distinguished traffic in one direction from that going in the reverse direction[17]. These models also allowed for 'call-back' effects, in which calls in one direction affected the volume of calling in the reverse direction. Among other things, these models can be viewed as the initial attempts to give coherence to the call externalities described in Section 3.

As before, the concern is with two areas, A and B, that are not part of the same local-calling area. Let Q_{AB} denote calls from A to B, Y_A income in A, P_{AB} the price from A to B, T_A the number of telephones in A, and T_B the number of telephones in B. With this notation, the model of the preceding section becomes

$$\ln Q_{AB} = \alpha_0 + \alpha_1 \ln Y_A + \alpha_2 \ln P_{AB} + \alpha_3 \ln(T_A \cdot T_B). \tag{11}$$

We now make two modifications to this model. The first is to allow for telephones in the sending area and receiving areas to have different elasticities, while the second modification is to allow for a 'reverse-traffic' effect, whereby calling from A to B is affected by the volume of traffic from B to A. Expression (11) accordingly becomes

$$\ln Q_{AB} = \alpha_0 + \alpha_1 \ln Y_A + \alpha_2 \ln P_{AB} + \alpha_3 \ln T_A + \alpha_4 \ln T_B + \alpha_5 \ln Q_{BA}. \tag{12}$$

Because the reverse-traffic effect is reciprocal, there will also be an equation for the traffic from B to A:

$$\ln Q_{BA} = \beta_o + \beta_1 \ln Y_B + \beta_2 \ln P_{BA} + \beta_3 \ln T_B + \beta_4 \ln T_A + \beta_5 \ln Q_{AB}. \tag{13}$$

These two equations form a simultaneous system and must be estimated jointly.

[16] Pacey (1983) was the first attempt to model toll demand on a route-specific basis.

[17] The first modeling efforts taking reciprocal calling and call-back effects into account were undertaken at Southwestern Bell in the mid-1980s. The results are reported in Larson, Lehman, and Weisman (1990). This study will be discussed in Section 5.

5. Empirical studies

Let me now turn to the empirical literature on telecommunications demand. As noted in the introduction, deregulation and competition have had mixed effects on telecommunications demand research. While the literature has continued to grow, its nature has changed and additions to it are now less informative about the empirical structure of the industry than was the case prior to the breakup of AT&T in 1984. Price elasticities are now viewed as important trade secrets, and few studies become openly available for rapidly growing new products and services, such as cellular and Internet access. As a consequence, empirical information about the structure of telecommunications demand is becoming increasingly dated and incomplete. With this in mind, my approach in this section will be to review a handful of studies that, in my view, ably summarize the present state-of-the-art in telecommunications demand modeling. In doing this, the focus will be on primarily on methodology, rather than empirical results, *per se*. Seven key studies will be reviewed, three for residential access demand and four for toll demand.

5.1. Studies of residential access demand and local usage; residential access demand: Taylor and Kridel, 1990

As noted in the preceding section, the first study that made use of the discrete-choice/consumer's surplus framework in analyzing the household demand for access to the telephone network was Perl (1983). Perl's model was estimated using data from the Public Use Sample of the 1980 Census of the Population. Data on the presence/absence of telephone service, income, and a variety of socio-demographic variables from the Census sample were combined with telephone rates, at a wire-center level, from records of the local Bell Operating Companies. To forestall identification of any individual household, the Census Bureau discloses place of residence only for areas of at least 100,000 population. This creates obvious difficulties in matching rates with households, for in virtually all cases areas of 100,000 or more contain several wire centers. The consequences of this may be of little importance when the focus is national, but problems clearly emerge when attention is shifted to specific socio-demographic groups in sparsely settled regions and areas. These considerations led Southwestern Bell to develop an alternative to Perl's model in which the unit of observation is a census tract, rather than an individual household. The census tract is the smallest geographical area that can be analyzed with census data, and provides for a much better match between rates and place of residence than is possible using data for individual households[18].

[18] A census tract typically consists of about 600 households. Although aggregation obviously leads to a loss of information, tracts are sufficiently diverse in income levels and socio-demographic characteristics that the parameters of most interest can continue to be isolated.

We can pick up the Southwestern Bell model at the second line of the probability statement for a household demanding access in expression (5) above, namely,

$$P(\ln CS > \ln \pi) = P(u > \ln \pi - a + \alpha p - \beta \ln y). \tag{14}$$

If it is assumed that u is N(0, σ^2), the probability of access will accordingly be given by

$$P(access) = 1 - \Phi(\ln \pi - a + \alpha p - \beta \ln y), \tag{15}$$

where Φ denotes the distribution function for the normal distribution.

With census tracts as the unit of observation, the inequality,

$$u > \ln \pi - a + \alpha p - \beta \ln y, \tag{16}$$

must be aggregated over the households in a tract. To do this, it is assumed that income, y, is log-normally distributed, independent of u, with mean μ_y, and variance σ_y^2. Let

$$v = u + \beta \ln y. \tag{17}$$

It then follows that v will be N(μ_y, $\sigma^2 + \beta^2\sigma_y^2$). Let P_j denote the proportion of households in census tract j that have a telephone. This proportion will be given by

$$P_j = P(v_j > \ln \pi_j - a + \alpha p_j), \tag{18}$$

or equivalently,

$$\begin{aligned} P_j &= P(w_j > w_j^*) \\ &= 1 - \Phi(w_j^*), \end{aligned} \tag{19}$$

where

$$w_j = \frac{v - \beta\mu_{yj}}{[\sigma^2 + \beta^2\sigma_{yj}^2]^{1/2}}. \tag{20}$$

and

$$w_j^* = \frac{\ln\pi_j - a + \alpha p_j - \beta\mu_{yj}}{[\sigma^2 + \beta^2\sigma_{yj}^2]^{1/2}}. \tag{21}$$

Let F_j denote the probit for census tract j calculated from expression (19), and write

$$F_j = \frac{\ln\pi_j - a + \alpha p_j - \beta\mu_{yj}}{[\sigma^2 + \beta^2\sigma_{yj}^2]^{1/2}} + \varepsilon_j, \tag{22}$$

where ε is a random error term. Clearing the fraction, we have

$$F_j(\sigma^2 + \beta^2\sigma_{yj}^2)^{1/2} = \ln\ \pi_j - a + \alpha p_j - \beta\ \mu_j + \varepsilon_j^*, \tag{23}$$

where

$$\varepsilon_j^* = (\sigma^2 + \beta^2\sigma_{yj}^2)^{1/2}\ \varepsilon_j. \tag{24}$$

Equation (23) offers a number of challenges for estimation, but before discussing these the fact that customers in some areas can choose between either flat-rate or measured service must be taken into account. Three possibilities must be considered:

(1) Only flat-rate service is available;
(2) Only measured service is available;
(3) Both flat-rate and measured service are available.

In flat-rate only areas, access will be demanded if [from expression (16), since $p = 0$]

$$u\ >\ \ln\ \pi_f - a - \beta\ \ln y, \tag{25}$$

while in measured-rate only areas, access will be demanded if

$$u\ >\ \ln\ \pi_m - a + \alpha p - \beta\ \ln y, \tag{26}$$

where π_f and π_m denote the monthly fixed charges for flat-rate and measured services, respectively. In areas where both flat-rate and measured service are options, access will demanded if *either* (25) or (26) hold. From these two inequalities, it follows that access will be demanded if

$$u\ >\ \min(\ln\ \pi_f - a - \beta\ln y,\ \ln\ \pi_m - a + \alpha p - \beta\ln y). \tag{27}$$

Finally, given that access is demanded, flat-rate service will be selected if

$$\ln\ \pi_f\ <\ \ln\ \pi_m + \alpha p, \tag{28}$$

while measured service will be chosen if[19]

$$\ln \pi_m + \alpha > \ln \pi_f. \tag{29}$$

To incorporate these choices into the analysis, define the dummy variables:
$\delta_1 =$ 1 if only flat-rate service is available; 0 otherwise;
$\delta_2 =$ 1 if only measured service is available; 0 otherwise;
$\delta_3 =$ 1 if both flat-rate and measured service are available; 0 otherwise.

With these dummy variables, we can rewrite Equation (23) as

$$\begin{aligned} F_j(\sigma^2 + \beta^2\sigma_{yj}^2)^{1/2} = {} & \delta_1 \ln \pi_{fj} + \delta_2(\ln \pi_{mj} + \alpha\, p_j) \\ & + \delta_3 \min(\ln \pi_{fj},\ \ln \pi_m + \alpha\, p_j) - a - \beta\, \mu_j + \varepsilon_j^*. \end{aligned} \tag{30}$$

With the addition of a set of socio-demographic variables, this is the equation that was estimated by Taylor and Kridel (1990)[20]. The model was estimated from a data set consisting of 8,423 census tracts from the 1980 Census in the five states then served by Southwestern Bell, Arkansas, Kansas, Missouri, Oklahoma, and Texas.

Several things are to be noted regarding the estimation of this model:

(a) The equation is non-linear in β;
(b) σ^2, the variance of u is a parameter that has to be estimated;
(c) Min($\ln \pi_{fj}$, $ln\ \pi_m + \alpha\, p_j$) must be calculated for the areas in which both flat-rate and measured service are available. But to do this requires knowledge of α.

Estimation proceeds by first rewriting the model as

$$Z_j = -a - \beta\, \mu_j - \Sigma\gamma_i X_{ij} + \varepsilon_j^*, \tag{31}$$

where

$$Z_j = F_j(\sigma^2 + \beta^2\sigma_{yj}^2)^{1/2} - \delta_1 \ln \pi_{fj} - \delta_3 \min(\ln \pi_{fj},\ \ln \pi_{mj} + \alpha\, p_j) \tag{32}$$

and where X_{ij} denotes predictors other than income. As none of Southwestern Bell's territory at the time had mandatory measured service, δ_2 is dropped.

[19] While choice is deterministic (given α), bill minimization is *not* assumed. Flat-rate expenditures may exceed measured-rate expenditures, but inequality (26) could still hold. Rearranging the inequality in (28), flat-rate will be chosen over measured service if $\ln\pi_f < \ln\pi_m + \alpha p$, i.e., if the difference in the logarithms of the monthly fixed charges is less than the price elasticity of usage (under measured service).

[20] The socio-demographic characteristics represented included owner/renter, age, size of household, rural/urban, ethnicity (white, black, Hispanic, native American), number of years at address, and employment status. Local-loop mileage charges were also taken into account, as was the number of lines that could be reached in the local-calling area.

If σ^2, β,and α were known, z_j could be calculated (using observed values for F_j, σ^2_{yj}, π_{jj}, π_{mj}, and p_j) and equation (31) could be estimated as a linear regression. However, these parameters are not known, so estimation proceeds by an iterative search procedure as follows:

1. The initial search is over values of σ^2 for fixed values of β and α. The value of σ^2 that is finally selected is the one that maximizes the correlation between the actual and predicted values of F_j.
2. The second step is to fix σ^2 at the value obtained in step i, keep β fixed as in step i, and search over α. Again, the value of α selected is the one that maximizes the correlation between the actual and predicted values of F_j.
3. The final step involves iteration on β. An iteration consists of the estimation of equation (31), with z_j [from equation (32)] calculated using the values of σ^2 and α from steps 1and 2 and β from the immediately preceding iteration. Iterations proceed until the estimate of β stabilizes[21].

5.2. Bypass via EAS: Kridel (1988)

As is well known, the toll-to-local subsidy was justified historically by allocating a substantial portion of the costs of local plant to toll service. At the time of the AT&T divestiture, these non-traffic-sensitive costs totaled about $16 billion, of which about $9 billion was the interstate portion and about $7 billion the intrastate portion. With divestiture, most of these non-traffic-sensitive costs were to be recovered through traffic-sensitive per minute charges at both ends of a toll call. Since these charges were seen as 'payment' to the toll carrier's switch (or point-of-presence), those at the sending end could be avoided if the customer were to bypass the local exchange company by connecting directly with its toll carrier's point-of-presence. As bypass facilities are not inexpensive, for direct bypass to pay requires a large volume of toll calling, and accordingly is not feasible for the vast majority of residential customers. However, in a circumstance in which there is a large volume of toll calling between adjacent regions, merging the regions into a local-calling area through Extended Area Service (*EAS*) could effect bypass[22]. Bedroom communities adjacent to large central cities provide cases in point, and in 1987–1988 several of these in Texas were using the regulatory forum to demand lower toll rates via optional *EAS* calling plan.

In anticipation of building the necessary infrastructure for provision of the *EAS* calling plans, Kridel (1988) undertook a study at Southwestern Bell to assess the

[21] The estimation procedure was evaluated before applying it to the Southwestern Bell data set by a Monte Carlo study. The Monte Carlo results show that the estimators have a moderate bias that appears to diminish with sample size. The results also show that the variance of the error term u has little effect on the accuracy of the estimated coefficients. For details, see Taylor and Kridel (1990) or Kridel (1987)

[22] Other studies of the demand for Extended Area Service include Pavarini (1976, 1979) and Martins-Filho and Mayo (1993).

likely demand. The point of departure for the analysis is the standard utility maximization framework in which, given a choice between two calling plans (toll and *EAS*), the customer is assumed to choose the one that yields the greater utility (as measured by consumer surplus). Specifically, if the increase in consumer surplus associated with *EAS* is greater than the subscription price, the customer will choose *EAS*. The increase in consumer surplus consists of two parts: (1) the toll savings due to the decrease to zero in the price of calls, and (2) the benefits arising from additional calling. Let the former be denoted by *TS* and the later by V_s. *EAS* accordingly will be purchased if:

$$\begin{aligned} \Delta CS &= TS + V_s \\ &> p_{EAS} \end{aligned} \tag{33}$$

where ΔCS denotes the change in consumer surplus and p_{EAS} denotes the *EAS* subscription price.

Since ΔCS is not directly observable, the choice criterion in expression (33) is reformulated in probability terms as:

$$\begin{aligned} \Delta CS &= \Delta CS^* + u \\ &> P_{EAS}, \end{aligned} \tag{34}$$

where ΔCS^* represents the observed or deterministic portion of the increase in consumer surplus and u is a well-behaved error term. The probability of a customer choosing *EAS* will then be given by:

$$P(EAS) = P(u > \gamma(p_{EAS} - \Delta CS^*)) \tag{35}$$

where γ is a parameter than can be thought of as the inclination of a subscriber to take advantage of a given net benefit (or loss) resulting from the selection of *EAS*.

As no historical data were then available, the model was estimated using purchase intentions data obtained in a survey of customers in the areas affected. Customers were queried whether *EAS* would be purchased at a given subscription price, and this price was varied across respondents. The resulting price variation allowed for predictions to be made of take-rates at various subscription prices. The customer, when providing a rational response to the *EAS* purchase query, is (at least subconsciously) estimating the benefits to be derived from additional calling (V_s), for the customer was informed of his/her toll savings (*TS*) just prior to the purchase question. Since the benefit in question is directly related to the amount of call stimulation, it is natural to treat the customer's new usage level (q_N) as a parameter to be estimated.

For the choice model represented in expression (35) to be estimated, the term ΔCS^* must be specified in terms of observable quantities. Accordingly, Kridel assumed that the demand function for q has the 'Perl' form,

$$q = A\mathrm{e}^{-\alpha p} y^{\beta}, \tag{36}$$

where:

q = minutes of toll usage into the EAS area
p = toll price
y = income
A, α, β = parameters.

With this functional form, the consumer surplus under each option will be given by:

$$\begin{aligned} CS_0^* &= \int_{P_0}^{\infty} a\mathrm{e}^{-\alpha p} y^{\beta} \mathrm{d}p \\ &= q_0/\alpha \qquad\qquad \text{non-}EAS \end{aligned} \tag{37}$$

$$\begin{aligned} CS_N^* &= \int_{0}^{\infty} a\mathrm{e}^{-\alpha p} y^{\beta} \mathrm{d}p \\ &= q_N/\alpha \qquad\qquad EAS, \end{aligned} \tag{38}$$

where

p_0 = current toll price
q_0 = observed usage at p_0
q_N = the satiation level of usage with EAS (since $p_N = 0$).

From these two expressions, $\Delta\ CS^*$ will be given by

$$\begin{aligned} \Delta CS^* &= CS_N^* - CS_0^* \\ &= (q_N - q_0)/\alpha. \end{aligned} \tag{39}$$

While the satiation level of usage q_N is unobserved, it can be calculated from the demand function in (36) as

$$q_N = q_0 \mathrm{e}^{\alpha p_0}. \tag{40}$$

At this point, the choice model is seen to involve two parameters, α and γ. The parameter α is related to the price elasticity of demand for toll, namely, αp, while γ (as noted earlier) can be thought of as the inclination of the subscriber to take

advantage of a given net benefit resulting from selection of *EAS*. The higher (in absolute value) is γ, the more likely (everything else remaining constant) a customer will be to purchase *EAS*. Kridel extends the model to allow for individual customer effects by assuming the parameters α and γ to be functions of income and subscriber demographics. This allows for all customers to have the same form of demand function, but to have different price elasticities and inclinations to purchase *EAS*. Demand-type variables (income, household size, etc.) are included in α, while choice-influencing variables (perceptions, education, etc.) are included in γ.

The stochastic term u in expression (35) is assumed to have a logistic distribution, thus implying a logit form for the choice model . Hence, the probability of a customer choosing *EAS* will be given by

$$P(u > \gamma(p_{EAS} - \Delta CS^*)) = \frac{1}{1 + \exp(\gamma x p_{EAS} - \Delta CS^*))} \tag{41}$$

The model was estimated for a sample of 840 households drawn at random from suburban communities using a hybrid Berndt-Hausman-Hall-Hall (1975) and quasi-Newton optimization routine.

Once estimated, the model was used to predict the take-rate and usage stimulation corresponding to *EAS* subscription prices of \$5, \$20, and \$25 per month. At a subscription price of \$5, the model predicted a take-rate of nearly 76% and an increase of nearly 93% in usage. At a price of \$25, the predicted take-rate is a little over 50% and an increase in usage of about 65%. The interesting and (and significant) thing about these stimulation estimates is that they were about twice as large as would have been generated using Southwestern Bell's (then) existing price elasticities for toll calling[23].

5.3. Choice of class of local service: Train, McFadden, and Ben Akiva (1987)

The third 'local' study to be discussed is a study of the demand for local service, by Train, McFadden, and Ben Akiva (TMB [1987]) that was undertaken in the mid-1980s by Cambridge Systematics, Inc. for a large East Coast local exchange company. Unlike the Taylor-Kridel study, where the choice is access/no access, this study focuses on the choice of class of service, given that some form of access is already demanded. What sets the TMB study apart is that the class of service is determined jointly with a portfolio of calls. Calls are distinguished according to

[23] This should not be taken to mean that the existing elasticities were wrong, but only that they were not applicable to the question that was being asked. One can reasonably infer that the households whose data were being analyzed exhibited above-average price responses because they were drawn from suburban communities that were already petitioning for EAS. The price elasticities obtained by Kridel were applicable to such circumstances, but almost certainly not to short-haul toll calls in general. More will be said about this in Section 6.

number, duration, distance, and time-of-day, and a portfolio of calls are accordingly seen as consisting of a particular number of calls of a particular duration to a set of specific locations at specific time of the day. Households in the data set analyzed faced a variety of service options, ranging from 'budget-measured' to 'metropolitan-wide flat-rate' service, so that the cost of a portfolio varies with the particular service option selected.

TMB employ a conditional (i.e., nested) logit framework in which the choice of service is assumed to be conditional on the portfolio of calls that is selected. With this structure, the probability of a household choosing a particular service option can be interpreted as depending upon the household's expected portfolio of calls (reflecting, for example, the tendency of households who place a lot of local calls to choose flat-rate service). Since usage is in turn conditional on access, however, the portfolio that a household will actually choose in a month will depend upon the service option that has been selected, as this determines the cost per call. In the data set analyzed, three service options were available to all households, and two additional services were available to some households. Portfolios were defined for 21 time zones. The probability of observing a particular (s, i) service-portfolio combination, given available service and portfolio options, is assumed to be nested logit. In the nesting, alternatives [where an alternative is an (s, i) combination] are nested together with the same portfolio, but different service options[24].

5.4. *Studies of toll demand, point-to-point toll demand: Larson, Lehman, and Weisman (1990)*

Larson, Lehman, and Weisman (1990) were the first to formulate a model incorporating reverse (or reciprocal) calling. They estimated their model using city-pair data from Southwestern Bell's service territory. The data used in estimation were quarterly time series for the period 1977 Quarter 1 through 1983 Quarter 3. Nine city pairs were analyzed. The city pairs were defined *a priori*, and for the most part were selected on a natural community-of-interest based on nearness. The larger cities were designated as points A and the smaller cities as points B. The nine city pairs thus form 18 routes to be analyzed, 9 AB routes and 9 BA routes.

[24] TMB provide an excellent discussion of the considerations that determine an appropriate nesting. The key consideration is the pattern of correlations among the omitted factors. For local calling, most of the specific factors that determine portfolio choice–such as where the household's friends and relatives live, the time and location of activities that require use of the telephone, etc.–are not observed. These factors, however, are similar across all service offerings for any portfolio. This leads to nesting together alternatives with the same portfolio, but different service options. On the other hand, the primary factor affecting the choice of service is price, which is observable. (The price of a portfolio under a particular service option is simply the cost of the portfolio under the rate schedule associated with that option.) Since alternatives with the same service option, but not the same portfolio of calls, can be similar with respect to observed factors, but not (at least relatively) with respect to unobserved factors, these are not nested.

The two-equation model estimated by LLW was as follows:

$$\ln Q_{it} = \alpha_0 + \sum_{j=1}^{g} \alpha_{ii}\delta_{ij} + \sum_{j=1}^{h} \beta_j \ln P_{it-j} + \sum_{j=1}^{l} \gamma_j \ln Y_{t-j} + \lambda \ln POP + \xi \ln Q_{it}^{*} + u_{it} \tag{43}$$

$$\ln Q_{it}^{*} = a_0 + \sum_{j=1}^{g} a_i\delta_{ij} + \sum_{j=1}^{h} b_j \ln P_{it-j} + \sum_{j=1}^{l} C_j \ln Y_{t-j} + \mathrm{d}\ln POP + \mathrm{e}\ln Q_{it} + v_{it} \tag{44}$$

where:

Q_{it} = Total minutes of intraLATA long-distance traffic between the ith city pair. Traffic refers to DDD (direct-distance-dialed) calls only, aggregated over all rate periods, that originate in city A and terminate in city B. Q_{it}^{*} is defined similarly, and represents the reverse traffic from B to A.

Pi = Real price of an intraLATA toll call on the ith route, defined as a fixed-weight Laspeyres price index, expressed as average revenue per minute, and deflated by the appropriate state or SMSA CPI.

Y_i = Real per-capita personal income in originating city.

POP = Market size, defined as the product of the populations in cities A and B.

$\delta_{ij} = 1$ for $i = j$, 0 otherwise.

The models were estimated with pooled data for the nine city pairs by two-stage least squares. The error terms u_i and v_i were assumed to be homoscedastic within a route, but heteroscedastic across routes. Contemporaneous covariance across routes was allowed for, as was first-order autocorrelation within routes. Finally, it was assumed that $\mathrm{E}(u_{it}v_{jt'}) = 0$ for all i,j,t, and t'. Estimation proceeded as follows[25]: To begin with, reduced-form equations were estimated using least-squares-dummy-variables (LSDV), corrected for autocorrelation and heteroscedasticity and tested for functional form using the Box-Cox format. Fitted values from these equations were then used in direct 2SLS estimation of the structural equations. The 2SLS residuals were tested for autocorrelation and heteroscedasticity, and the structural equations were re-estimated with appropriate corrections for autocorrelation and heteroscedasticity, and then tested for correct functional form. Final estimation of the structural equations was performed by LSDV (using the initial fitted values for the endogenous variables) with an addition correction for contemporaneous covariances (across routes, but not across equations).

[25] Full details of the estimation are given in Larson (1988).

The LLW was a pioneering effort, and is accordingly subject to the usual problems of something being cut from whole cloth with only a vague pattern for guidance. The use of per-capita income, while the dependent variable is aggregate and not per-capita, is questionable, and almost certainly contaminates the elasticities for market size, and probably for income as well. There may also be a problem in constraining the elasticities for the sizes of the two calling areas to be equal. Such a constraint makes sense when calling between the two areas is aggregated, but probably not when the model is directional, and feedbacks from reverse traffic are allowed for. Traffic from the sending area will continue to depend upon the number of telephones in that area, but the reverse-traffic effect will probably swamp the effect of the number of telephones in the receiving area, if for no other reason than reverse traffic is a better proxy for community of interest. Finally, use of the population as a proxy for the number of telephones may be a problem as well.

These criticisms, however, in no way undermine LLW's central result, namely, that calls in one direction tend to generate calls in return that are independent of price and income. This result, in the LLW models, is statistically strong, and has been confirmed in subsequent studies at Bell Canada and elsewhere[26]. The reverse-calling phenomenon is clearly most relevant in situations in which the same telephone company does not serve both directions of a route. In this situation, a decrease in the price on the BA route will lead to the stimulation of the traffic on the AB route in the absence of a price change on that route. Failure to take this into account can lead to serious biases in the estimates of price elasticities[27].

5.5. *Toll demand models estimated from a sample of residential telephone bills: Rappoport and Taylor (1997)*

As noted in the introduction, one of the costs of competition and deregulation has been a severe narrowing of data available for the estimation of telecommunications demand models. Even if access to proprietary data were not a problem, fragmentation is. With multiple carriers, data from an individual carrier is restricted to that carrier's customers. And while local exchange companies in the U.S. still do much of the billing for long-distance carriers, the only information that in general can be made available by them to outsiders are the records of their own customers. To get a complete picture of the telecommunications consumption of a household or business, one must go to the household or business directly. The need for doing this is increasingly being recognized, and telecommunications data bases now exist that are derived from information obtained directly from the

[26] See Appelbe et al. (1988) and Acton and Vogelsang (1992).

[27] In particular, one must be careful to distinguish between uni-directional elasticities and bi-directional elasticities. Uni-directional elasticities allow for a change in price in one direction only, while bi-directional elasticities allow for prices changes in both directions. In both cases, calculations must be made from the reduced-forms of the models. For the formulae involved, see Appelbe et al. (1988).

records of users. We now turn to a recent study by Rappoport and Taylor (1997), which one of the first to exploit such data.

Rappoport and Taylor estimate models for four categories of residential toll calling in the U.S.: intraLATA, interLATA/intrastate, interLATA/interstate, and total interLATA using data from telephone bills obtained from a sample of about 9,000 U.S. households[28]. Besides the novelty of the data set used in estimation, the study offers several additional innovations, including the use of a call concentration index that is constructed from the call detail that is available in the individual telephone bills. For the interLATA categories, the primary interexchange carrier is taken into account, as well as whether the household subscribes to an optional calling plan. Finally, the models estimated allow for cross-price effects among the different toll categories.

The basic model relates the aggregate number of minutes of toll calling for a residential customer to price, income, and a variety of socio-demographic characteristics, including race (white, black, Hispanic), education, and size of the urban area in which the household resides. Price is measured as average revenue per minute, calculated as the total bill for the toll category in question divided by the total number of billed minutes for the category[29]. Income, as well as all of the socio-demographic variables, is measured in terms of a vector of dummy variables, corresponding to income categories. The call concentration index is defined as the percentage of a household's calls in a toll category that go to three or fewer numbers.

[28] The data used in estimation are from the first Bill Harvesting™ survey of PNR & Associates (now TNS Telecoms) of Jenkintown, PA. The data in Bill Harvesting are obtained from residential telephone bills purchased from households in nationally representative samples of U.S. households. The sample in the first survey consisted of about 12,000 households, of which about 9000 provided the information requested. The PNR surveys are ongoing, and now include businesses as well as households. Additional information can be obtained from TNS Telecoms at http://www.PNR.com.

[29] The use of average revenue per minute as a measure of price is controversial in the literature, and has been criticized by myself (Taylor [1980, 1994]), as well as others. The problem, as is well-known, is that, because of possible measurement error and the presence of multi-part tariffs, a dependence between average revenue per minute and the error term can come into play that leads to a biased and inconsistent estimate of the price elasticity. With the data being analyzed in this study, this is not thought to be a problem. The reason is that the primary determinant of the calls that a household makes is the household's 'community of interest.' The tariff structure can be a factor in determining the *volume* of calling that a household makes, but its community of interest is the primary factor in determining the *mix* of calling, in the sense of where and to whom calls are directed. While in general the community of interest is unobservable, this should not be of consequence so long as it is not itself dependent upon the tariff structure. With regard to this, it seems reasonable that households do not choose their communities of interest on the basis of the cost of calling into them. Since communities of interest obviously vary across households, there will accordingly be as many 'prices' – even with the same tariff structure – as there are separately identifiable communities of interest. On the other hand, this is not to say that the potentiality of non-independence should be ignored, but rather should be tested for as part of the analysis. If independence is rejected, then estimation will need to take it into account. The ideal, of course, is to use prices that are derived from the actual underlying tariff structure, but this is a counsel of perfection. Failing this, the standard is to employ an instrumental variable estimator, as in the study of Taylor (1996) discussed below.

The index is intended as a measure of a household's community of interest, the interpretation being that the higher is the index, the more concentrated is the household's community of interest. The concentration index is included in the model both directly and interacted with price. The final variables included in the model are dummy variables for AT&T, MCI, and Sprint referring to the household's choice of primary (i.e., 1-plus) interexchange carrier and a dummy variable indicating whether or not the household subscribes to an optional calling plan[30].

Since not all households make toll calls, the analysis proceeds in two stages. In the first stage, a discrete-choice probit model is estimated that explains the probability that a household has a non-zero amount of toll calling. A Mills ratio is calculated from this probit equation and its inverse is then used as an independent variable in a second-stage model that explains the demand for toll calling for those households with non-zero calling activity[31]. The base sample underlying the analysis consisted of 6,461 households, and was the sample used in estimating the first-stage probit models from which Mills ratios were calculated. The second-stage models (i.e., the toll-usage models) were then estimated from sub-samples of households whose total month's toll activity was between 4 and 500 minutes and whose average toll cost per minute was less than 35 cents[32]. Sample sizes range from about 2,450 households for the interLATA/intrastate category to about 4,600 for interLATA/interstate and for interLATA/interstate combine.

The results yielded an estimated price elasticity for intraLATA toll of about -0.44 and an elasticity for interstate toll of about -0.50. On the whole, different categories of toll calling are found to be substitutes, although the effects are indicated to be quite small. Call concentration is found to have a strong inverse effect on the size of the price elasticity, with low concentration being associated with a large (in absolute value) elasticity and vice versa. With regard to inter-exchange carrier effects, AT&T customers, on the average, have lower toll activity than non-AT&T customers, but for those households who subscribe to an optional calling plan, AT&T customers have higher calling volume than non-AT&T customers.

[30] Toll minutes and price are household-specific as obtained from the household's telephone bill. Income and the socio-demographic variables are measured as means for the ZIP+4 postal areas in which a household resides.

[31] The Mills ratio (or more precisely, the inverse of the Mills ratio) adjusts for the mean of the error term in the second-stage model not necessarily being equal to zero as a consequence of the second-stage sample including only those households with non-zero toll-calling activity. For a discussion of how the Mills ratio arises and how it is calculated, see Greene (1997).

[32] The subsamples were truncated at the upper end of toll usage in order to eliminate toll activity that was almost certainly primarily business motivated, and at the lower end in order to eliminate households whose toll calling was unusually infrequent or was low during the month of the sample because of travel or unavoidable absence from home. Also, some households who subscribe to an optional calling plan had (for whatever reason) little or no toll activity during the month of the sample. Because of the monthly subscription fee, the average prices per minute will be unduly high in these cases, and their presence would clearly bias the estimation. Restricting the sub-samples to households having a calculated average price per minute of less than 35 cents eliminates the distortion that these artificially inflated average prices per minute would cause.

5.6. *Competitive own- and cross-price elasticities in the intraLATA toll market: Taylor (1996)*

In the mid-1990s, there was a rush to open intraLATA toll markets to carriers other than the incumbent local exchange telephone companies. As competition emerges in what had previously been a monopoly market, a variety of new elasticities comes into play. A single industry demand function is replaced by demand functions faced by individual firms, whose elasticities will differ from the industry elasticity both because of cross-price effects among firms and because of heterogeneity across consumers. Whereas a monopoly supplier may treat consumers as homogeneous, this will almost certainly not be the case with multiple suppliers, as firms seek to establish individual 'market niches' based upon socio-demographic characteristics and differing price elasticities among subgroups of consumers. I want now to describe briefly a study that I undertook in 1996 which attempts to isolate some of these intraLATA competitive effects. The data set employed was from the second wave of the Bill Harvesting surveys of PNR & Associates. Information was available in the second Bill Harvesting survey on both the total amount of residential intraLATA toll traffic, as well as a breakdown of this traffic between that carried by the incumbent local exchange company and that handled by other carriers. When this information is combined with toll rates in each of the states for the local exchange company and AT&T (taken as a surrogate for the other intraLATA carriers), both own- and cross-price elasticities can be estimated.

The basic model used in estimation is similar to the one used by Rappoport and Taylor in the study just described. The aggregate number of minutes of intraLATA toll calling for a residential customer is related to price, income, age, an index of call concentration and a dummy variable for whether or not the household subscribes to an optional calling plan[33]. Competition in the intraLATA market is allowed for through the inclusion of two price variables as predictors and a competition dummy variable. The first price variable is the price for the incumbent local exchange company (LEC), while the second price variable is an AT&T[34]. Competition is also allowed for through a dummy variable that is defined in terms of whether or not a household used a non-LEC carrier in its intraLATA calling. This dummy variable is included both by itself and interacted with the LEC price as well as the AT&T price. The expectation, of course, is that

[33] At the start of the analysis, a wide variety of socio-demographic variables were also considered (including age, ethnicity, size of metropolitan area resided in, and education). However, the price elasticities are little affected by the presence or absence of these variables, and age is the only one retained in the models that are tabulated.

[34] The AT&T price data refer to 4-minute calls for day, evening, and night based upon tariffs in effect in January 1995. Aggregate means for a household have been obtained by weighting the day, evening, and night prices by the corresponding numbers of intralata calls as recorded in the household's bill. Unlike in the Rappoport-Taylor study, an instrumental-variable estimator is used for the LEC price, obtained by regressing average revenue per minute on a set of state dummy variables.

the price elasticity will be larger when competition is present. The competition dummy variable is also interacted with the dummy variables indicating the presence or absence of subscription to optional intraLATA and interLATA calling plans. Finally, call concentration is measured as the percentage of a household's intraLATA calls that are made to four or fewer numbers[35]. The index is included in the model both linearly and interacted with prices, which allows the price elasticities to vary with the level of concentration.

Once again, the analysis proceeds in two stages. In the first stage, a probit model is estimated which explains the probability that a household has nonzero toll calling. A Mills ratio is calculated from this probit equation and its inverse is used as an independent variable in a second-stage model to explain the demand for toll calling for those households with nonzero toll activity. As before, the sample used in estimated is restricted to households with intraLATA toll minutes between 4 and 500 (or alternatively, between 4 and 1000).

In brief, what the empirical results show is as follows:

(1) In general, competition clearly shifts and alters the structure of intraLATA demand.
(2) For households who do all of their intraLATA calling through their local exchange company, the estimated price elasticity is -0.24.
(3) When a competitive carrier is used as well as the local exchange company, the elasticity increases (in absolute value) to -0.41, with a 'cross' elasticity of 0.23. A 'net elasticity of -0.18 is thereby implied.

At this point, there must be caution as to how these elasticities are interpreted. Do the demand functions that are estimated represent demand functions for households, or are they to be interpreted as the demand functions facing the local exchange company? Since the data are for individual households, it is clear that the vantage point must be that of the *customers* of intraLATA carriers, rather than of that of the *carriers* themselves. If it is assumed that all households in the sample had access to a non-LEC carrier (even if by no means other than through 1-0-XXX), then the LEC-only elasticity can be interpreted as the elasticity which pertains to the households that remain 'loyal' to the LEC. By the same token, the 'net' elasticity, i.e., the difference between the LEC and AT&T elasticities, can be interpreted as the elasticity pertaining to households who use multiple carriers[36].

[35] Empirically, it makes little difference whether three or four calls form the cut-off for measuring call concentration.

[36] The only problem with this interpretation is that, since 'competition' in the analysis is defined ex post in terms of households which actually use a non-LEC carrier (although not necessarily exclusively), the elasticities must accordingly be interpreted as conditional on a household being 'loyal' or 'non-loyal.' In other words, there is no allowance for a price change leading to a 'loyal' household becoming 'non-loyal, or vice versa. To allow for this requires the specification and estimation of a model to describe the probability of movement from one group to the other. See Kridel, Rappoport, and Taylor (1998), also Tardiff (1995).

5.7. *Two-stage budgeting and the total bill effect in toll demand: Zona and Jacob (1990)*

I now turn to what has been an often-recurring question in analyzing telecommunications demand, whether households approach their expenditures for telecommunications services in terms of a total bill. The usual assumption in demand analysis – and the one that has been implicit so far in this discussion – is that telecommunications services – local, toll, features, etc. – compete along with all other goods and services in a household's market basket for the household's income. In this framework, short-haul toll calls can be either a substitute or a complement for long-haul calls, and an increase in access charges need not impinge upon either. An alternative view is that households budget for a total amount to be spent on telecommunications services, so that an increase in access charges, say, will necessarily lead to less being spent on usage. At one level, the debate can be cast in terms of what is the appropriate budget constraint to be used in telecommunications demand functions: is it income or a budgeted total expenditure for telecommunications? Should the household's preferences be appropriately separable (whether weakly or strongly need not concern us), the two approaches are of course equivalent, for the optimization problem can be formulated in terms of a two-stage budgeting. In the first stage, the household decides how to allocate its income between telecommunications and all other goods and services, while in the second stage it allocates the total telecommunications expenditure determined in the first stage among the different telecommunications services (including access).

Whether or not a household's preferences are appropriately separable is ultimately an empirical question, and has received little systematic attention. The one place that I am aware where it has is an unpublished paper by Zona and Jacob (1990). The motivation for the Zona-Jacob paper was to test a hypothesis, even stronger than separability, that had been current for some time, namely, the view that all that matters is the total bill[37]. Households are assumed to alter their demands in response to changes in the telecommunications prices in order to keep the total bill unchanged. The practical difference between this assumption and separability (or two-stage budgeting) is in the total-bill response to price changes in individual telecommunications services. With two-stage budgeting, a household will reassess its total phone bill in response to a price change, whereas with the other assumption, the total bill will not be reassessed, but demand will be altered so that total spending remains the same.

The Total-Bill hypothesis was tested by Zona and Jacob using the Almost Ideal Demand System (AIDS) of Deaton and Muellbauer (1980). The AIDS specification is a budget share model, which relates expenditure shares to prices and the level of total expenditure. The AIDS system involves flexible functional forms and

[37] See Doherty and Oscar (1977).

does not limit substitution, but it easily allows in estimation for the restrictions across elasticities that are imposed by the two hypotheses. The Zona-Jacob procedure is to estimate three models, an unrestricted model, in which the budget constraint is income and no restrictions imposed on the cross elasticities, and two restricted models corresponding to the Two-Stage-Budgeting and Total-Bill hypotheses, respectively. The unrestricted model has four share equations, for each of interLATA toll, intraLATA toll, local service (access, custom-calling features, inside wiring, etc.), and all other goods. For good i, the AIDS specification is given by

$$w_i = \alpha_i + \sum_j \gamma_{ij} \ln p_j + \beta_i \ln(x/P), \tag{45}$$

where w_i denotes the budget p_j is the price of good j, x is income, and P is the quadratic price deflator defined by Deaton and Muellbauer (1980). Certain restrictions are imposed on the parameters across equations in order to ensure consistency (adding-up, symmetry, etc.) with standard demand theory[38]. Under Two-Stage-Budgeting, the second-stage conditional budget share for good i will be given by

$$w_i^* = \alpha_i^* + \sum_j \gamma_{ij} \ln p_j + \beta_i^* \ln(B/P^*), \tag{46}$$

where w_i^* is the bill share for service, B is the total telecommunications bill, and P^* is a suitably defined quadratic price deflator.

In the first stage, households are assumed to allocate income between telecommunications and all other goods. Zona and Jacob again assume an AIDS specification,

$$w^G = \alpha + \gamma \ln(P_B/CPI) + \beta \ln(x/P^{**}), \tag{47}$$

where w^G is the telecommunications budget share, P_B is the telecommunications group price index, CPI is the price of other goods, and P^{**} is once again a suitably defined quadratic deflator.

The data used by Zona and Jacob in estimation were from four distinct areas served by the (then) United Telecommunications System[39]. Each area was in a

[38] See Deaton and Muellbauer (1980, p. 316).

[39] Since United Telecommunications was not part of the Bell System, it could continue to offer both local and long-distance service (both intra- and interlata) after the AT&T divestiture. The Zona-Jacob study was accordingly of great deal of interest, not only because of its focus on the Bill-Targeting hypothesis, but also because of the uniqueness of its data set. GTE was in similar circumstances as United in providing both local and long-distance services, but I am not aware that a Zona-Jacob type study has ever been undertaken using GTE data.

different state and had different tariffs. The time frame for the data was June-December 1989. The telephone data were aggregated to the three services, local, intraLATA toll, and interLATA toll. The variables developed included the average total bill (total revenue divided by estimates of the number of households in the areas), bill shares for the three services, price indices for the three services (defined as average revenue per unit), a state-specific price index, and estimates of average household income. The unit of observation was a month, so that the data set consisted of 28 observations (7 months, 4 areas). A fixed-effects (i.e., least-squares-dummy-variable) pooled time-series/cross-section framework was used in estimation.

The results obtained by Zona and Jacob show little difference in the own- and cross-price elasticities for inter- and intraLATA toll for the unrestricted and the two-stage budgeting models, but a big difference between these two models and the bill-targeting model. Lagrange multiplier tests fail to detect a difference between the unrestricted and the two-stage budgeting models, but reject the bill-targeting model in favor of two-stage budgeting. The results accordingly support two-stage budgeting, but not bill-targeting.

6. An overview of what we know about telecommunications demand

At the time of my 1980 book, we knew a lot more about the demand for usage than the demand for access, a lot more about toll usage than local usage, a lot more about residential demand than business demand, and we probably had more precise knowledge about price elasticities than income elasticities. There was some knowledge of international demand, a little about business toll demand, even less about business WATS (i.e., 800 service) and private-line demands, and virtually nothing about business demand for terminal equipment. Also, virtually nothing was known regarding cross-elasticities between local and toll usage or between other categories of service. We knew that network and call externalities were important conceptual elements of telecommunications demand, but there was little empirical information as to their magnitudes. Option demand was also known to be a potentially important component of access demand, but again virtually nothing was known of its empirical importance. A continuum of price elasticities had emerged, with access at the bottom (in terms of absolute value) and long-haul toll and international calls at the top. Price elasticities clearly appeared to increase with length-of-haul. There was some evidence of elastic demand in the longest-haul toll and international markets, but there seemed little question that the toll market as a whole was inelastic. Income elasticities, for the most part, appeared to be in excess of 1.

During the 1980s and 1990s research on telecommunications demand has:

1. Added greatly to our knowledge of the demand for residential access. The small, but nevertheless non-zero, access elasticities obtained by Perl (1978)

have been confirmed and sharpened in the studies of Cain and MacDonald (1991), Kaserman, Mayo, and Flynn (1990), Perl (1983), and Taylor and Kridel (1990) for the U.S., J. P. Bodnar et al. (1988) and Solvason (1997) for Canada, and Rodriguez-Andres and Perez-Amaral (1998) for Spain.

2. Made a significant start in quantifying the structure of business demand, especially with regard to standard access and the demands for key, PBX, and centrex systems. The studies of Ben-Akiva and Gershenfeld (1989) and Watters and Roberson (1991) particularly come to mind[40].
3. Confirmed and sharpened previous estimates of price elasticities for medium- to long-haul toll usage. Relevant studies include Gatto, Kelejian, and Stephan (1988), Larson, Lehman, and Weisman (1990), Zona and Jacob (1990), the Bell-Intra models of Bell Canada ((1984, 1989), and the Telecom Canada models (Appelbe et al. (1988)).
4. Added to and greatly sharpened previous estimates of the price elasticities for short-haul toll. The studies that contributed to this include Kling and Van der Ploeg (1990), Kridel (1988), Martins-Filho and Mayo (1993), Train, McFadden, and Ben-Akiva (1987), Train, Ben-Akiva, and Atherton (1989), Watters and Roberson (1992), Zona and Jacob (1990).
5. Made an important start on obtaining estimates of cross-price elasticities between measured- and flat-rate local service, local service and toll, local usage and toll, and key, PBX, and centrex systems. Studies to consult include Ben-Akiva and Gershenfeld (1989), Kling and Van der Ploeg (1990), Kridel (1988), Train, McFadden, and Ben-Akiva (1987), Train, Ben-Akiva, and Atherton (1989), Watters and Roberson (1991), Zona and Jacob (1990).
6. Made a start (Kridel and Taylor [1993]) in analyzing the demand for custom-calling features, both singly and in bundles.
7. Made a start (Ahn and Lee [1999]) in analyzing the demand for access to cellular and wireless networks.
8. Made a start in analyzing the demand for access to the Internet (Rappoport et al. [1998], Kridel, Rappoport, and Taylor [1999, 2000]).

On the modeling front, three innovations stand out in the literature of the 1980s and 1990s: (1) the widespread use of quantal-choice models, (2) the analysis of toll demand on a point-to-point basis, and (3) application of random-coefficients methodology to a system of toll demand equations. With the much greater emphasis on access demand in the 1980s it is natural that quantal-choice models would find much greater application. In the pre-1980 literature, the only study

[40] One might argue that customer-premise equipment should be approached in a technology-adoption framework, whereby various types of network and 'lock-in' effects are taken into account (Cf., Katz and Shapiro [1986]). This makes clear and obvious sense when the focus is on the demand for customer-premise equipment, *per se*, and also when the focus is on the demand for *high-speed* access and usage. Even so, it needs to be kept in mind that the demand for access and customer-premise equipment, in the end, always flows from the benefits from usage, rather than the other way around.

that used a quantal-choice framework was Perl (1978). In the years since, studies using this framework included Ben-Akiva and Gershenfeld, Kling and Van der Ploeg (1990), Kridel (1988), Kridel and Taylor (1993), Perl (1983), Taylor and Kridel (1990), Train, McFadden, and Ben-Akiva (1987), Train, Ben-Akiva, and Atherton (1989), J. P. Bodnar et al. (1988), and Solvason (1997). The most sophisticated application is by Train et al. where (as seen in Section 5 above) a nested logit model is used to analyze the choice among standard flat-rate, EAS, and measured services and the choice of particular portfolio of calls by time-of-day and length-of-haul.

As was noted in Section 5, important impetus to modeling toll demand on a point-to-point basis came in 1985 with the Larson-Lehman-Weisman study at Southwestern Bell using city-pair data. While there had been earlier point-to-point studies [including Larsen and McCleary (1970), Deschamps (1974), and Pacey (1983)], the LLW study was the first to allow for reverse-traffic effects. Acton and Vogelsang (1992) have subsequently applied this framework to inter-provincial toll demand in Canada by Telecom Canada (1988) and to international calling. While both the LLW and Telecom Canada studies employed pooled time-series/cross-section data sets in estimation, the most sophisticated application of varying- and random-coefficients methodology is that of Gatto, Kelejian, and Stephan (1988). Interstate toll demand is approached in a multi-equation framework in which toll demand is disaggregated by mileage bands, time-of-day, and non-operator/operator handled calls. The model is estimated in a random-coefficients, variance-components format using a pooled data set consisting of quarterly observations for the 48 contiguous U.S. states and the District of Columbia[41].

Progress has also been made in clarifying understanding of the network and call externalities. In great part, this occurred as a result of protracted debate before the Canadian Radio-television and Telecommunications Commission in the mid- to late-1980s concerning the size of the coefficients for market size in the Bell Canada's toll demand models[42]. While precise estimates of the magnitudes of the externalities are unfortunately still lacking, the relevant questions are now sharper. In 1980, the biggest concern regarding consumption externalities was whether the network externality might be large enough to justify continuing the toll-to-local subsidy. Little attention was given to the call externality because it was thought that, even if one of any magnitude should exist, it would be internalized between calling parties through call-back effects. What such a view overlooked is

[41] Two interesting new concepts were introduced into applied demand analysis in Gatto-Kelejian-Stephan study, namely, stochastic symmetry and stochastic weak separability. Stochastic symmetry is taken to mean that the Slutsky symmetry conditions across commodities in a cross section are only satisfied stochastically, while stochastic weak separability means that the price elasticities in a cross section are assumed to be stochastically proportional to the income elasticity.

[42] See Bell Canada (1984, 1989), Breslaw (1985), Globerman (1988), and Taylor (1984). See also de Fontenay and Lee (1983), and Taylor (1994, Appendix 4).

the likelihood that calls give rise to further calls, quite independently of price and income. Strong empirical evidence that such is the case is given in the point-to-point toll demand models of Southwestern Bell and Telecom Canada, both of which indicate that a call in one direction stimulates something like one-half to two-thirds of a call in return. The call externality thus seems to have a dynamic dimension that goes beyond a simple cost sharing of calls between two parties. As noted in Section 3, it is best, in my view, not to attribute all of the call-back effect to call externalities, but to see part of it as reflecting an independent phenomenon, or what I have called the *dynamics of information exchange*.

In light of all that has gone on in the telecommunications industry since 1980, an obvious question to ask is whether the changes and upheavals that have taken place have caused similar upheavals in the structures of telecommunications demands[43]. In my view, there is little evidence that this has been the case, at least for the large categories like residential and business access and medium- to long-haul toll. I find no evidence that the toll market as a whole has become either more or less elastic than in 1980[44]. Residential access price elasticities, in contrast, appear to have become smaller[45], but this would seem to be a consequence of higher residential penetration rates[46].

[43] This is for the market as a whole. Price elasticities for individual interexchange carriers are another matter, as are also elasticities for certain submarkets within a given market. Because of a decrease in market share, the price elasticities that AT&T faces are clearly larger now than (say) in 1990. For MCI and Sprint, however, they are almost certainly smaller. For attempts to estimate firm-specific demand elasticities, see Ward (1994), Kahai, Kaserman, and Mayo (1996), and Kaserman et al. (1999).

[44] One of the more interesting results resulting from research during the 1980s relates to the price elasticity for short-haul toll calls of (say) 15 to 40 miles. In 1980, the view was that this elasticity was quite small, of the order (say) of -0.2. There was no real evidence in support of this number; it simply fit into the flow of elasticities as one went from shorter to longer lengths-of-haul. The evidence now points to a value of this elasticity as being in the range of -0.25 to -0.40. (See Duncan and Perry [1995], Hausman [1991], Kling and Van der Ploeg [1990], Kridel [1988], Martins-Filho and Mayo [1993], Rappoport and Taylor [1997], Train, McFadden, and Ben-Akiva [1987], Train, Ben-Akiva, and Atherton [1989], and Zona and Jacob [1990]). Nevertheless, it should be emphasized that this should not be viewed as an instance of structural change, but rather the emergence, for the first time, of credible evidence as what the value of this elasticity actually is. For an account of the consequences of the failure on the part of the Regulatory Authority in California to take this value seriously, see Hausman and Tardiff (1999).

[45] Cf., however, Cain and MacDonald (1991), who obtain residential access price elasticities that are about double Perl's, especially at very low income levels. As their results are based upon data from the mid-1980s, Cain and MacDonald interpret these higher elasticities as reflecting structural change. Whether this represents genuine structural change, or is simply a reflection of sharp shifts in the ethnic and social composition of households, though, is not clear.

[46] Price elasticities calculated from the probit/logit models that are used in estimating access demand vary inversely with the level of penetration. A point that is worthy of mention is that, in comparing the results using data from the 1970 Census with those with data from the 1980 Census, Perl (1983) found an upward shift in the willingness-to-pay by households to have telephone service that could not be attributed to income or to changes in socio-demographic factors. It would be interesting to know whether data from the 1990 Census would show the same phenomenon *vis-...-vis* 1980. However, I am not aware that such a study has been undertaken.

In all candidness, however, the real concern at the moment with structural change in telecommunications is not with traditional categories of wireline access and usage, but with new forms of access and usage that are associated with technological change and the emergence of the Internet[47]. The list includes:

1. A worldwide explosion in cellular/mobile telephony. Is wireless telephony a substitute for traditional wireline communication or a complement? What are the price and income elasticities for wireless? Has the widespread availability of wireless access increased the total market for telecommunications? These are questions for which there are virtually no answers at the present, at least in the open literature[48].
2. The demand for access to the Internet. Among other things, this has led to a sharp acceleration in the demand for second lines[49].
3. The emergence of Internet telephony. The question here is not whether, but how and how much, traditional circuit-switched telephony will be affected. At the present Internet telephony is essentially a 'new' service, for which there is no empirical information at all about its likely formal structure. Indeed, one of the biggest challenges facing the empirical analysis of Internet telephony is how to go about collecting relevant data to begin with.
4. The demand for broadband access to the Internet. There are two questions related to broadband access: (1) what will be the size of the market, and (2) which technology is going to win out on the supply side – cable, traditional wireline (via DSL or similar technology), or wireless[50]?

Returning to more traditional concerns, the post-1980 literature confirms the finding that the absolute value of the price elasticity increases with the length-of-haul. Indeed, all considered, this probably remains the best-established empirical regularity in telecommunications demand. Interestingly enough, however, no convincing explanation as to why this ought to be the case has ever been offered. Some have suggested that it reflects primarily a price phenomenon: that since price elasticity (for non-constant elasticity demand functions) in general vary with

[47] For an informed discussion of some of the problems involved, see Cracknell and Mason (1999).

[48] The 'closed' literature is obviously another matter, for it is precisely this type of information that has become closely guarded proprietary assets. At the annual International Communications Forecasting Conference in San Francisco in June 1998, two analysts from SBC discussed a study of cellular demand using data of Cellular One. Their presentation was methodologically extremely informative, but devoid of numerical information about elasticities. In a quick perusal of the recent literature, I find only one econometric study for wireless demand, Ahn and Lee (1999).

[49] For recent studies of the demand for additional lines, see Cassel (1999), Duffy-Deno (2001) and Eisner and Waldon (2001). For a pre-Internet study of the demand for second lines in Canada, see Solvason and Bodnar (1995) and Solvason (1997). For an initial effort to estimate a price elasticity for Internet access, see Kridel, Rappoport, and Taylor (1999).

[50] The only broadband access studies that I am presently aware of are Madden, Savage, and Coble-Neal (1999), Madden and Simpson (1997), Eisner and Waldon (2001), and Kridel, Rappoport, and Taylor (2001).

the level of price, elasticities for longer-haul calls will be larger simply because they cost more[51]. On the other hand, an argument can also be made that the phenomenon represents a genuine distance effect, in that (for whatever reason) calls with a longer-haul are simply more price-elastic.We now turn to factors that might account for this relationship. Unfortunately, surprisingly little is known about either the characteristic of motivations associated with calls to different distances. The only study that I am aware of that provides useful information in this regard is from a household survey in Los Angeles and Orange Counties in California that was undertaken by GTE of California in the mid-1980s[52]. From the 680 households that participated in the survey, information was collected from a log of calls made by the households during a two-week period. The information collected included date, time of call, duration of call, destination, person making the call, the party called, purpose of call, and importance of call. Summary physical characteristics of 22,351 calls that were analyzed include:

1. A high degree of skewness in the distribution of distance; half of all calls terminated within six miles of the calling party.
2. A high degree of skewness in distribution of duration; half of all calls lasted four minutes or less.
3. A significant day-to-day variation of calls:
   - Calls more frequent on weekdays than on weekends;
   - Calls on weekdays less frequently to longer distances than is the case on weekends;
   - Calls on weekdays shorter in duration than is the case on weekends (mean 6.1 minutes vs. 7.2 minutes).
4. A significant variation in the rate period of calls:
   - 48 percent of calls during the business day; remainder nearly evenly split between evening (28 percent) and night/weekend (24 percent);
   - Mean call length-of-haul steadily increases from business day (33.4 miles) to evening (67.6 miles) to night/weekend (105.2 miles);
   - Mean call duration increases from business day (5.6 minutes) to over 7 minutes in evening and night/weekend periods;
   - Mean cost of calls increases from business day ($0.361) to night/weekend ($0.446) to evening ($0.496).

[51] This follows from the definition of price elasticity as $(\partial_q/\partial_p)(p/q)$, where q and p denote quantity and price, respectively. Strong evidence against the price-elasticity/length-of-haul relationship being just a price phenomenon is provided by results from the study of Gatto et al. (1988) for interstate toll traffic in the U.S. The price elasticity obtained in that model, -0.72, agrees closely with pre-1980 price elasticities for long-haul calls (see Taylor [1994, Appendix 1]), but it is estimated from a data set in which the real price of toll calling was sharply lower (because of competition) than the levels of the 1970s. Also, it should be noted that the emergence of 'postalized' pricing, in which toll rates do not vary by distance, pretty much makes the price/distance relationship a moot point.

[52] See Mead (1985).

Summary attitudinal characteristics of calls include:

1. Significant variation in call purposes:
   - Greatest proportion of calls (33.8 percent) are to friends, followed by calls to relatives (26.6 percent);
   - Calls to friends are to shorter distances, similar duration (8.1 minutes), less costly, and of lesser importance than calls to relatives;
   - Work-related calls are less frequent (10.5 percent), shorter, less costly, and more important than social calls.
2. Differentiation of calls by callers:
   - Two-thirds of calls are by wives;
   - Calls by teenage girls are longest (mean 9.1 minutes vs. overall mean of 6.7 minutes);
   - Husbands are more likely to make calls to work; children's calls are predominately to friends.

With the foregoing physical and attitudinal characteristics of calling behavior as the background, I believe that it is now possible to put together a convincing story that can account for the observed positive relationship between the size of toll price elasticities and length-of-haul[53]. The place to begin is to remind us of an often-forgotten point: that price elasticity consists of two components, an income effect and a substitution effect. The substitution effect is measure of the extent to which goods or services can substitute for one another when there is a price change without making the consumer any worse off in terms of economic welfare. The income effect, on the other hand, is a measure of the extent to which the consumer's real income is changed as a result of the price change. Ordinarily, the importance of the income effect is represented by the importance in the consumer's budget of the good whose price has changed. Goods whose expenditures account for a small proportion of the consumer's total expenditure will have a small (or even tiny) income effect, while a good whose expenditures account for a large proportion of total expenditure will have a large income effect. Goods that are ordinarily viewed as necessities (such as heating fuel and telephone service) will also have relatively larger income effects the lower the level of income.

In assessing income effects, however, a point that is usually overlooked is the effect on the consumer's welfare of not consuming a good because of a price increase. In the case of making or not making a phone call because the call has become more expensive, the question that needs to be asked is what the consequences are (not necessarily in monetary terms) of not making the call. For residential consumers, this cost is usually cast in terms of the utility (or satisfaction) that is foregone by the call not being made. For many calls, however, this is not the correct measure of cost, for the call may be important to the

[53] The argument that follows is taken from Chapter 11 of my 1994 book.

earning of income, rather than simply being a way of *consuming* it. In this case, the actual income effect of not making a telephone call may be large, although the decrease in real income (as customarily measured) caused by the price increase may be extremely small. Note that what is at issue here is a negative income effect that arises from a call not being made, in contrast with the small negative income effect that arises from the actual making of the now more expensive call.

Let us now tie this consideration in with the communities of interest that are associated with calls of different lengths-of-haul. Although length-of-haul can obviously vary continuously, assume for simplicity that there are only four lengths: local calls, short-haul toll calls, medium-haul toll calls, and long-haul toll calls. These categories can be identified with local calling, intraLATA toll, intrastate toll, and interstate toll. In terms of numbers, the communities of interest of most households will be largest for the local calling area and then decrease progressively with length-of-haul. The results from the GTE survey (as well as common sense) suggest that the community of interest for local calls (again for most households) will consist of friends and relatives and people and acquaintances associated with work, shopping, and recreation. Friends and relatives become proportionately more important as the communities of interest become increasingly farther away. Interestingly, the GTE survey suggests that friends are relatively more important than relatives in the community of interest for local and short-haul calls, while the reverse is true for medium- and long-haul calls.

What, then, are the implications of this changing composition in the community of interest for the relative sizes of price elasticities? To begin with, it seems clear that calls that are related to work, shopping, and commercial recreation will be less sensitive to price than calls that are to friends and relatives, the major reason being the contingency noted above regarding the cost of not making a call. Almost certainly, this cost (whether psychic or monetary) will be higher for a call related to work, shopping, or commercial recreation, than a call to a friend or relative. Calls of the former type are simply of greater urgency (in the sense that their primary purpose is information exchange, rather than chatting and recreation), and there may be no feasible substitute for a telephone call (such as waiting for a friend or relative to call instead). For shopping and commercial recreation, the price of the call is probably treated as part of the price of the good or activity in question, in which case the price of the call will not be of much importance. This is a strong argument as to why calls of less than six miles are most numerous, shortest in duration, and have the smallest price elasticity of demand.

As we move to longer-haul calls, the relative importance of these two generic types of calls changes, as the communities of interest come to consist more and more of friends and relatives, and less and less of people associated with work, shopping, and commercial recreation. For short-haul toll calls, these latter

characteristics will still be a factor, but not to the extent as for local calls. Since the opportunity cost of not making a call will in general be of less consequence, calls and calling will become more sensitive to price, which in turn implies a larger price elasticity. The GTE survey shows that the calls in the mileage bands constituting the medium- to long-haul toll categories are predominantly to relatives. These calls tend to be concentrated in the evening and night/weekend time periods, and are significantly longer in duration than local or short-haul toll calls. In general, urgency is less critical, and substitute forms of communication are much more feasible. Also, for someone with several relatives or friends in a distant city or locale, a lengthy call to one can serve to communicate with all via local calls being made by the one who was called to begin with. All of these factors auger for a greater sensitivity of the volume of toll usage to price for longer-haul calls, and thus to a larger price elasticity.

Let me now turn to two areas for which extensive structural information is still lacking. As in 1980, business demand continues to head the list. Ben-Akiva and Gershenfeld (1989) made an important start in modeling the business access/equipment nexus of decisions, but their analysis was restricted to data for the public switched network. Griffin and Egan (1985) and Watters and Roberson (1991) represent efforts to estimate own- and cross-price elasticities among MTS, WATS, and private line. However, the Griffin-Egan study is based upon aggregate data, focuses only on usage, and provides no information on how the demands for MTS, WATS, and private line interact with the demand for terminal equipment or vary with size, type, or organizational structure of a business. The Watters-Roberson analysis, on the other hand, is restricted to data of Southwestern Bell, and hence represents a study of business demand as seen by a major telecommunications supplier, rather than the full and final demands of business users.

Finally, there is international demand. While a number of interesting and useful studies of international demand were undertaken in the 1980s and 1990s it nevertheless remains an under-researched area. Studies of note include Telecom Canada (Appelbe et al. [1988]), Acton and Vogelsang (1992), Cracknell (1999), Garin (1999), Garin and Perez (1996, 1998, 2000), and Hackl and Westlund (1995, 1996). The price elasticities in the Telecom Canada study are (as noted in Section 5) based on a point-to-point model and are reported only on a uni-directional basis (i.e., for a change in Canada-U.S. rates only). The elasticities are estimated to be of the order of -0.5, which are consistent with bi-directional Canada-U.S. elasticities of -0.9 or larger. The Acton-Vogelsang study is based upon annual data for minutes of calling between the U.S. and 17 Western European countries from 1979 to 1986. The volume of calling both originating and terminating in the U.S. was examined as a function of the price in the sending country, price in the receiving country, together with economic and conditioning variables such the volume of trade between the countries, the structure of employment, the number of telephones, and the price of telex services. Models are estimated for 1979–1982

and 1983–1986. Estimates of the uni-directional elasticities vary from insignificant to near -1.0[54].

The most comprehensive analysis of international calling for any single country in the literature is provided in the studies of Garin and Perez for Spain. In their 1996 paper, Garin and Perez analyze the demand for total outgoing international telephone calling from Spain using a time-series/cross-section model that is estimated with an annual data set for the years 1985–1989 for the 50 Spanish provinces. Models are estimated in which both real expenditure per line and the number of calls per line are related to real GDP, price, the number of overnight stays in Spanish hotels by foreign residents, and the number of foreign residents living in Spain. In their 1998 paper, Garin and Perez use a reciprocal-calling framework based on the Larson-Lehman-Weisman model to analyze international traffic between Spain and 27 African and Oriental countries. A similar framework is used by Garin (1999) to analyze calling traffic between Spain and 19 American countries and by Garin and Perez (1999) to analyze traffic between Spain and 24 other European countries. In these models, the number of outgoing minutes per line to country i is related to the volume of trade between Spain and country i, price, the number of tourists visiting Spain from country i, and the number of legal foreign residents from country $i$that reside in Spain. The number of incoming minutes is also included as a measure of reciprocal calling. The price elasticities obtained range from -0.5 to -0.85 to a high of -1.3 for the (bi-directional) elasticity for calls to Africa and the Orient. Reciprocal-calling effects are strong in all three studies.

The focus in the paper of Hackl and Westlund (1996) is on the demand for telecommunications between Sweden and three destination countries: Germany, the United Kingdom, and the United States. A model for each destination is estimated that relates outgoing traffic (measured in minutes) from Sweden to indices of industrial production, volume of trade, and price. A key concern of the paper is whether price elasticities vary over time, and two models are estimated that allows for this to occur, the first a Kalman filter and the second a moving local regression. In both cases, the results support price elasticities that increases (in absolute value) for calls to Germany and the United Kingdom and a price elasticity that decreases for calls to the United States[55].

[54] The models specified by Acton and Vogelsang allow for arbitrage and call externalities, but not for reverse-traffic effects of the type analyzed in the Southwestern Bell and Telecom Canada studies. Also, the models include only the number of telephones in the European countries. For a tabulation of studies of international demand, together with estimated price and income elasticities, that were undertaken prior to 1980, see Taylor (1994, Appendix 2).

[55] Unlike the other studies described in these paragraphs, Cracknell (1999) does not present any new empirical results, but provides, rather, an informative discussion of changes in the international market over the last 30 years and how these changes relate to changes in the size of international price elasticities.

7. Challenges for the future

Competition, deregulation, and new services made available by the convergence of computers, wireless, cable, and the Internet with conventional wire-line telephony, have made telecommunications demand analyses of the traditional types essentially exercises of the past. Even if it were desired to do, comprehensive industry studies of the type undertaken by Gatto, Kelejian, and Stephan (1988) or Gatto et al. (1988) are impossible to pursue, for it is no longer feasible to organize the required data bases. The distinction between access and usage is still the place to begin in studying telecommunications demand, but the purview of telecommunications demand has clearly gotten much larger. No longer is it adequate to look simply at wire-line access, intraLATA, interLATA, and international toll calling, 'vertical services,' and customer-premise equipment. There are now many other forms of types of access to consider, including broadband access to the Internet[56], and a variety of new forms of telephony, including wireless, fixed-wireless, and Internet telephony. Indeed, in my view, the Internet itself is now part and parcel of the telecommunications sector, and should be encompassed in its framework[57].

The theoretical models described in Section 4, and as applied in the studies described in Sections 5 and 6, still provide cogent points of departure. Indeed, this is a principal reason for their presentation in detail. But their application must now be to services and markets that reflect present telecommunications structures for suppliers as well as for users. Clearly, one of the biggest current challenges facing telecommunications demand modelers is to identify structures that can reasonably be assumed to be stable. Almost certainly, access to the telecommunications network can still be assumed to be stable, but most probably not with respect to the traditional definition of access in terms of fixed wire-line access. A broader definition of access is required[58]. Similar considerations apply to usage. While it is probably still reasonable to attribute stability to tele-

[56] A concern among many is whether broadband markets are going to bifurcate into 'haves' and 'have-nots' with regard to high-speed access (i.e., the so-called 'digital divide'). For the most part, this concern relates to possible extensions of the meaning of 'universal service', and is accordingly a policy and social welfare question. Nevertheless, demand analysis has an important role to play in assessing willingness-to-pay for those who, in the absence of subsidies, would be likely to be included amongst the 'have-nots'. How to model this in the absence of real behavorial data is, of course, a major obstacle, and will almost certainly have to entail the use of contra-factual survey data, such as, for example, could be obtained in the TNS Request™ surveys (See www.pnr.com.).

[57] As consumers, households as well as business, are increasingly demanding multiple types of access, access demand needs to be approached in terms of a portfolio of access 'pipes,' as opposed to just a single 'pipe.' A good first initial framework to employ would be the portfolio-choice model of Train, McFadden, and Ben-Akiva (1987).

[58] The same is also true for usage. Serious thought needs to given to measuring telecommunications traffic in terms of bandwidth, as opposed to simply minutes of usage, in order to accommodate data and broadband demands by both businesses and households.

communications at different distances, times-of-day, and day-of-week, this is probably not the case with respect to conventional definitions of intraLATA and interLATA fixed-line toll calling. Definitions must be expanded to include e-mail and other forms of Internet communication, as well as wireless-to-wireless connections[59].

Also, as has been emphasized in this chapter, there are major challenges of data. Traditional sources for telecommunications data clearly no longer suffice, or indeed are even relevant. Increasingly, data are going to have to be obtained directly from telecommunications users, rather than suppliers, or from companies whose businesses are simply to monitor and count.

Finally, one of the failings of the traditional corporate decision-making structure in the telephone industry has been that demand analysis has always played a spot role. It has been used and useful for calculating repression/stimulation effects of proposed tariff changes in regulatory proceedings, but never employed in a systematic way in budgeting, forecasting, and marketing[60]. Its home in the old Bell System was always the Revenue Requirements organization. As deregulation and competition have forced telephone companies to slim down, demand analysis has tended to be treated as unnecessary regulatory overhead. In truth, some slimming-down was probably in order. To allow demand analysis to languish, however, because of deregulation and competition, as has been done in U.S. and Canadian telecommunication companies, is a mistake. One of the biggest current challenges for demand analysts in the telecommunications companies is to forge links with marketing departments, and to become integrated in the company budgeting and forecasting processes. Applied demand analysis has a strong strategic role to play in a competitive environment, ranging from conventional types of elasticity estimation in traditional markets to the identification of new markets[61]. The possibilities are vast. It only requires imagination, hard work – and some humility!–on the part of economists and econometricians.

[59] In my view, the ultimate place to seek stability in household consumption patterns is in broad consumption categories. As suggested in Taylor (1998), I have to believe that there is basic stability in the structure defining expenditure for telecommunications in relation to other expenditure categories, as well as for functional categories of expenditure within telecommunications. The challenge, of course, is to define the functional categories in meaningful ways.

[60] The reference here is to U.S. telephone companies. Demand analysis and forecasting for years was well integrated into the corporate decision-making structure at Bell Canada. Indeed, from the mid-1970s through the mid-1990s, Bell Canada was one of the world's major centers of innovative telecommunications demand analysis activity. (Dineen and Abrar [1999] provide a summary and overview of the research that was done at Bell Canada during these years.) Now, however, virtually none of this activity remains.

[61] For an innovative framework for assessing the value of new services, see Hausman and Tardiff (1997). For overviews and frameworks for forecasting the demands for new products, see Palombini and Sapio (1999), Saether (1999), and Tauschner (1999).

References

Acton, J. P., and I. Vogelsang, 1992, Telephone demand over the Atlantic: Evidence from country-pair data, Journal of Industrial Economics, 40, 305–323.

Ahn, H., and M. H. Lee, 1999, An econometric analysis of the demand for access to mobile telephone networks, Information Economics and Policy, 11, 297–306.

Appelbe, T. W., N. A. Snihur, C. Dineen, D. Farnes, and R. Giorano, 1988, Point-to-point modeling: An application to Canada-Canada and Canada-U.S. Long-distance calling, Information Economics and Policy, 3, 358–378.

Artle, R., and C. Averous, 1973, The telephone system as a public good: Static and dynamic aspects, Bell Journal of Economics and Management Science, 4, 89–100.

Bell Canada, 1984, Econometric models of demand for selected intra-Bell long-distance and local services, Attachment 1 to Response to Interrogatory Bell (CNCP) 20 February 84-509 IC, Supplemental, July 1984, CRTC, Ottawa, Ontario.

Bell Canada, 1989, Intra-Bell customer-dialed peak and off-peak model updates, Attachment 1, CRTC Public Notice 1988-45; Bell Canada and British Columbia Telephone-Company Review of Methodologies Used to Model Price Elasticities.

Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hausman, 1974, Estimation and inference in nonlinear structural models, Annals of Economic and Social Measurement, 635–665.

Ben-Akiva, M., and S. Gershenfeld, 1989, Analysis of business establishment choice of telephone system, Cambridge, MA: Cambridge Systematics, Inc.

Bodnar, J., P. Dilworth, and S. Iacono, 1988, Cross-section analysis of residential telephone subscription in Canada, Information Economics and Policy, 3, 311–331.

Breslaw, J. A., 1985, Network externalities and the demand for residence long distance telephone service, Working Paper No. 1985-13, Concordia University, Montreal, Quebec.

Breslaw, J. A., 1989, Bell Canada and British Columbia telephone company: Review of methodology used to model price elasticity, Evidence on behalf of the Government of Quebec before the Canadian Radio-television and Telecommunications Commission, CRTC Telecom 1988-45, Ottawa, Ontario.

Cain, P., and J. M. MacDonald, 1991, Telephone pricing structures: The effects on universal service, Journal of Regulatory Economics, 3, 293–308.

Cassel, C. A., 1999, Demand for and use of additional lines by residential customers, in D. G. Loomis and L. D. Taylor, eds, The Future of the Telecommunications Industry: Forecasting and Demand Analysis, Boston: Kluwer Academic Publishers.

Cracknell, D., 1999, The changing market for inland and international calls, in D. G. Loomis and L. D. Taylor, eds, The Future of the Telecommunications Industry: Forecasting and Demand Analysis, eds, Boston: Kluwer Academic Publishers.

Cracknell, D., and C. Mason, 1999, Forecasting telephony demand against a background of major structural change, in D. G. Loomis and L. D. Taylor, eds, The Future of the Telecommunications Industry: Forecasting and Demand Analysis, Boston: Kluwer Academic Publishers.

Deaton, A. S., and J. Muellbauer, 1980, An almost ideal demand system, American Economic Review, 70, 312–326.

de Fontenay, A., and J. T. Lee, 1983, BC/Alberta long distance calling, in L. Courville, A. de Fontenay, and R. Dobell, eds., Economic Analysis of Telecommunications: Theory and Applications, Amsterdam: North-Holland.

Deschamps, P. J., 1974, The demand for telephone calls in Belgium, presented at Birmingham International Conference in Telecommunications Economics, Birmingham, England, Catholic University of Louvain, Louvain, Belgium.

Doherty, A. N., and G. M. Oscar, 1977, Will the rates produce the revenues? Public Utilities Fortnightly, 99, 15–23.

Duffy-Deno, K., 1999, Demand for additional telephone lines: An empirical note, Information Economics and Policy, 13, 283–299.

Duncan, G. M., and D. M. Perry, 1995, IntraLATA toll demand modeling: A dynamic analysis of revenue and usage data, Information Economics and Policy, 6, 163–178.
Eisner, J., and T. Waldon, 1999, The demand for bandwidth: Second telephone lines and on-line services, Information Economics and Policy, 13, 301–309.
Erikson, R. C., D. L. Kaserman, and J. W. Mayo, 1998, Targeted and untargeted subsidy schemes: Evidence from post-divestiture efforts to promote universal service, Journal of Law and Economics, 41, 477–502.
Garin-Munoz, T., 1999, Telephone traffic to America, Departamento Analysis Economico, Universidad Nacional de Educacion a Distancia, Madrid, Spain.
Garin, T., and T. Perez, 1996, Demand for international telephone traffic in Spain: An economic study using provincial panel data, Information Economics and Policy, 8, 289–315.
Garin, T., and T. Perez, 1998, Econometric modeling of Spanish very long distance international calling, Information Economics and Policy, 10, 237–252.
Garin-Munoz, T., and T. Perez-Amaral, 1999, A model of Spain-Europe telecommunications, Applied Economics, 31, 989–997.
Gatto, J. P., H. H. Kelejian, and S. W. Stephan, 1988, Stochastic generalizations of demand systems with an application to telecommunications, Information Economics and Policy, 3, 283–310.
Gatto, J. P., J. Langin-Hooper, P. Robinson, and H. Tryan, 1988, Interstate switched access demand, Information Economics and Policy, 3, 333–358.
Globerman, S., 1988, Elasticity of demand for long-distance telephone service, Report prepared by Steve Globerman Associates, Ltd. for the Federal-Provincial-Territorial Task Force on Telecommunications, Ottawa, Ontario.
Greene, W. H., 1997, Econometric Analysis, Englewood Cliffs, NJ: Prentice Hall Publishing Co, Third Edition.
Griffin, J. M., and B. L. Egan, 1985, Demand system estimation in the presence of multi-block tariffs, Review of Economics and Statistics, 67, 520–524.
Hackl, P., and A. H. Westlund, 1995, On price elasticities for international telecommunications demand, Information Economics and Policy, 7, 27–36.
Hackl, P., and A. H. Westlund, 1996, Demand for international telecommunication time-varying price elasticity, Journal of Econometrics, 70, 243–260.
Hausman, J. A., 1991, Phase II IRD testimony of Professor Jerry A. Hausman on behalf of pacific Bell Telephone Co., before the Public Utility Commission of California.
Hausman, J. A., T. J. Tardiff, and A. Bellinfonte, 1993, The effects of the breakup of AT&T on telephone penetration in the United States, American Economic Review, 83, 178–184.
Hausman, J. A., and T. L. Tardiff, 1997, Valuation of new services in telecommunications, in Dumont and J. Dryden, eds, The Economics of the Information Society, A. Luxembourg: Office for Official Publications of the European Communities, 76–80.
Hausman, J. A., and T. J. Tardiff, 1999, Development and application of long-distance price elasticities in California: A case study in telecommunications regulation, Cambridge, MA: National Economic Research Associates, Inc.
Kahai, S. K., D. L. Kaserman, and J. W. Mayo, 1996, Is the dominant firm dominant? An empirical analysis of AT&T's market power, Journal of Law and Economics, 39, 499–517.
Kahn, A. E., 1966, The tyranny of small business decisions: Market failure, imperfections, and the limits of economics, Kyklos, 19, 23–47.
Kaserman, D. L., J. W. Mayo, L. R. Blank, and S. K. Kahai, 1999, Open entry and local rates: The economics of IntraLATA toll competition, Review of Industrial Organization, 14, 303–319.
Kaserman, D. L., J. W. Mayo, and J. E. Flynn, 1990, Cross-subsidization in telecommunications: Beyond the universal service fairy tale, Journal of Regulatory Economics, 2, 231–250.
Katz, M. L., and C. Shapiro, 1986, Technology adoption in the presence of network externalities, Journal of Political Economy, 94, 822–841.

Kling, J. P., and S. S. Van Der Ploeg, 1990, Estimating local call elasticities with a model of stochastic class of service and usage choice, in A. de Fontenay, M. H. Shugard, and D. S. Sibley, eds, Telecommunications Demand Modeling, Amsterdam: North Holland.

Kridel, D. J., 1987, An analysis of the residential demand for access to the telephone network, Ph.D. Dissertation, Tucson, AZ: Department of Economics, University of Arizona.

Kridel, D. J., 1988, A consumer surplus approach to predicting extended area service (EAS) development and stimulation rates, Information Economics and Policy, 3, 379–390.

Kridel, D. J., Lehman, D. E., and D. L. Weisman, 1991, Options value, telecommunications demand and policy, St. Louis, MO: Southwestern Bell Telephone Co.

Kridel, D. J., P. N. Rappoport, and L. D. Taylor, 1998, Competition in intraLATA long-distance: Carrier choice models estimated from residential telephone bills, Information Economics and Policy, 13, 267–282.

Kridel, D. J., P. N. Rappoport, and L. D. Taylor, 1999, An econometric model of the demand for access to the Internet, in D.G. Loomis and L.D. Taylor, eds, The Future of the Telecommunications Industry: Forecasting and Demand Analysis, Boston: Kluwer Academic Publishers.

Kridel, D. J., P. N. Rappoport, and L. D. Taylor, 2001, An econometric model of the demand for access to the Internet by cable modem, in D.G. Loomis and L.D. Taylor, eds, Forecasting the Internet: Understanding the Explosive Growth of Data Communications, Boston: Kluwer Academic Publishers.

Kridel, D. J., and L. D. Taylor, 1993, The demand for commodity packages: The case of telephone custom calling features, Review of Economics & Statistics, 75, 362–367.

Larsen, W. A., and S. J. McCleary, 1970, Exploratory attempts to model point-to-point cross-sectional interstate telephone demand, Unpublished Bell Laboratories Memorandum, Murray Hill, New Jersey.

Larson, A. C., 1988, Specification error tests for pooled demand models, presented at Bell Communications Research/Bell Canada Industry Forum on Telecommunications Demand Analysis, Key Biscayne, FL, January 25–27, 1988, Southwestern Bell Telephone Co., St. Louis.

Larson, A. C., and D. E. Lehman, 1986, Asymmetric pricing and arbitrage, presented at the Sixth International Conference on Forecasting and Analysis for Business Planning in the Information Age, Southwestern Bell Telephone Co., St. Louis.

Larson, A. C., D. E. Lehman, and D. L. Weisman, 1990, A general theory of long-distance telephone demand, in Telecommunications Demand Modeling, A. de Fontenay, M. H. Shugard, and D. S. Sibley, eds, Amsterdam: North Holland Publishing Co.

Liebowitz, S., and S. Margolis, 2002, Network effects, Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science BV, 75–96.

Littlechild, S. C., 1975, Two-part tariffs and consumption externalities, Bell Journal of Economics, 6, 661–670.

Madden, G., S. J. Savage, and G. Coble-Neal, 1999, Subscriber churn in the Australian ISP market, Information Economics and Policy, 11, 195–208.

Madden, G., and M. Simpson, 1997, Residential broadband subscription demand: An econometric analysis of Australian choice experiment data, Applied Economics, 29, 1073–1078.

Martins-Filho, C., and J. W. Mayo, 1993, Demand and pricing of telecommunications services: Evidence and welfare implications, RAND Journal of Economics, 24, 439–454.

McFadden, D. L., 1978, Modeling the choice of residential location, in A. Karquist et al., eds, Spatial Interaction Theory and Planning Models, Amsterdam: North Holland Publishing Co.

Mead, J. E., 1985, Analysis of characteristics of residential telephone calling and propensities to alter calling activity per call, Thousands Oaks, CA: General Telephone Company of California.

Pacey, P. L., 1983, Long-distance demand: A point-to-point model, Southern Economic Journal, 49, 1094–1107.

Palombini, I. M., and B. Sapio, 1999, Analysis of customer expectations for the introduction of new telecommunications services, in D. G. Loomis and L. D. Taylor, eds, The Future of the

Telecommunications Industry: Forecasting and Demand Analysis, Boston: Kluwer Academic Publishers.
Pavarini, C., 1976, The effect of flat-to-measured rate conversions on the demand for local telephone usage, unpublished Bell Laboratories Memorandum, Bell Laboratories, Murray Hill, New Jersey.
Pavarini, C., 1979, The effect of flat-to-measured rate conversions on local telephone usage, in J.T. Wenders, ed, Pricing in Regulated Industries II, Denver, CO: Mountain States Telephone Co.
Perl, L. J., 1978, Economic and social determinants of residential demand for basic telephone service, White Plains, NY: National Economic Research Associates, Inc.
Perl, L. J., 1983, Residential demand for telephone service 1983, prepared for Central Services Organization of the Bell Operating Companies, White Plains, NY: National Economic Research Associates, Inc.
Rappoport, P. N., and L. D. Taylor, 1997, Toll price elasticities estimated from a sample of U.S. residential telephone bills, Information Economics and Policy, 9, 51–70.
Rappoport, P. N., L. D. Taylor, D. J. Kridel, and W. Serad, 1998, The demand for Internet and online access, in E. Bohlin and S. L. Levin, eds, Telecommunications Transformation: Technology, Strategy and Policy, Amsterdam: IOS Press.
Rodriguez-Andres, A., and T. Perez-Amaral, 1998, Demand for telephone lines and universal service in Spain, Information Economics and Policy, 10, 501–514.
Rohlfs, J., 1974, A theory of interdependent demand for a consumption service, Bell Journal of Economics and Management Science, 5, 16–37.
Saether, J. P., 1999, Limits to growth in telecommunications markets, in D. G. Loomis and L. D. Taylor, eds, The Future of the Telecommunications Industry: Forecasting and Demand Analysis, Boston: Kluwer Academic Publishers.
Solvason, D. L., 1997, Cross-sectional analysis of residential telephone subscription in Canada using 1994 data, Information Economics and Policy, 9, 241–264.
Solvason, D. L., and J. P. Bodnar, 1995, Cross-sectional analysis of subscription to additional residential telephone lines in Canada using 1992 data, Hull, Quebec: Bell Canada.
Squire, L., 1973, Some aspects of optimal pricing for telecommunications, Bell Journal of Economics and Management Science, 4, 525–525.
Tauschner, A., 1999, Forecasting new telecommunication services at a 'Pre-Development' product stage, in D. G. Loomis and L. D. Taylor, eds, The Future of the Telecommunications Industry: Forecasting and Demand Analysis, Boston: Kluwer Academic Publishers.
Tardiff, T. J., 1995, Effects of presubscription and other attributes on long-distance carrier choice, Information Economics and Policy, 7, 353–366.
Taylor, L. D., 1980, Telecommunications demand: A survey and critique, Cambridge, MA: Ballinger Publishing Co.
Taylor, L. D., 1984, Evidence of Lester D. Taylor on behalf of Bell Canada concerning the price elasticity of demand for message toll service in Ontario and Quebec, Ottawa, Ontario: Canadian Radio-television and Telecommunications Commission, November.
Taylor, L. D., 1994, Telecommunications demand in theory and practice, Boston: Kluwer Academic Publishers.
Taylor, L. D., 1996, Competitive own- and cross-price elasticities in the intraLATA toll market: Estimates from the Bill Harvesting II database, Tucson, AZ: Department of Economics, University of Arizona.
Taylor, L. D., 1998, Telecommunications infrastructure and economic development, in D. Orr and T. A. Wilson, eds, The Electronic Village: Policy Issues of the Information Economy, Ottawa, Ontario: C. D. Howe Institute.
Taylor, L. D., and D. J. Kridel, 1990, Residential demand for access to the telephone network, in A. de Fontenay, M. H. Shugard, and D.S. Sibley, eds, Telecommunications Demand Modeling, Amsterdam: North Holland Publishing Co.

Train, K. E., M. Ben-Akiva, and T. Atherton, 1989, Consumption patterns and self-selecting tariffs, Review of Economics & Statistics, 71, 62–73.

Train, K. E., D. L. McFadden, and M. Ben-Akiva, 1987, The demand for local telephone service: A fully discrete model of residential calling patterns and service choices, The RAND Journal of Economics, 18, 109–123.

Von Rabenau, B., and K. Stahl, 1974, Dynamic aspects of public goods: A further analysis of the telephone system, Bell Journal of Economics and Management Science, 5, 651–669.

Ward, M., 1994, Market power in long distance communications, Washington, D.C.: Federal Trade Commission.

Watters, J. S., and M. L. Roberson, 1991, An econometric model for dedicated telecommunications services: A systems approach, St. Louis, MO: Southwestern Bell Telephone Co.

Watters, J. S., and M. L. Roberson, 1992, An econometric model for IntraLATA MTS pricing plans, St. Louis, MO: Southwestern Bell Telephone Co.

Weisbrod, B. A., 1964, Collective-consumption of individual consumption goods, Quarterly Journal of Economics, 78, 471–477.

Weisman, D. L., 1988, Default capacity tariffs: Smoothing the transitional regulatory asymmetries in the telecommunications market, Yale Journal on Regulation, 149–178.

Zona, J. D., and R. Jacob, 1990, The total bill concept: Defining and testing alternative views, presented at BELLCORE/Bell Canada Industry Forum, Telecommunications Demand Analysis with Dynamic Regulation, Hilton Head, SC, April 22–25, Cambridge, MA: National Economic Research Associates.

Chapter 5

ECONOMETRIC COST FUNCTIONS

MELVYN A. FUSS

University of Toronto

LEONARD WAVERMAN

London Business School

Contents

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

The knowledge of costs is crucial in any industry, and has proven to be of special importance in telecommunications. The telecommunications industry has long been organized in most jurisdictions as a monopoly regulated by the state. Hence, the firm has had the knowledge of its costs, yet the state and the regulator have needed to regulate prices and to determine where entry, if any, should be allowed. The telecommunications carrier provides multiple services over a certain geographic area using capital that is fixed and common across a variety of applications. In addition, the impact of technological advance is also unevenly spread over the range of assets and services. Hence, a determination of costs, especially at an individual service level, is inherently difficult. Moreover, since the carrier, the regulator and the entrants have had different objectives, the information publicly available has always been limited. Academic economists have played important roles in the public policy process, and econometric cost functions were used more in this industry than in any other context.

As competition has grown, the debates concerning the presence or absence of natural monopoly elements in telecommunications have gradually subsided. The importance of public knowledge of firm-specific costs has not, however, correspondingly decreased. Where entrants compete with incumbents that originally were the sole providers of network services, the issues of the price (cost) to interconnect at, and of the wholesale price of 'unbundled' services have become crucial. Therefore, issues of 'correct social costs' have not diminished. What were, in the 1970's and 1980's, questions of natural monopoly, sustainable prices and unsustainable cross-subsidies have become questions of access pricing and entry conditions.

In this paper, we survey the uses and abuses of econometric cost function analysis in the telecommunications sector. Simultaneously, in the past decade, applications of accounting based techniques[1] and operations research based techniques[2] to study performance and the nature of production in the telecommunications sector have steadily increased. We do not survey these techniques in detail. We briefly highlight the role of operations research based techniques, but restrict ourselves to a more extensive discussion of econometric based techniques in this survey. Detailed and extensive reviews of the accounting and operations research based literatures are left as future agenda items.

1.1. Monopoly or competition

A hotly contested issue in the 1970's and 1980's, and in the early 1990's was whether competition in the provision of various services, public long distance or

[1] See Banker, Chang and Majumdar (1996) and Fraquelli and Vanoni (2000).
[2] See Majumdar (1995) and Uri (2001).

message toll (toll) services, local services and even in terminal equipment in the customer's home or office, was in the public interest. Incumbent telephone companies consistently argued that the provision of each and every one of these services was a natural monopoly. Hence, a competitive market structure was inefficient. On the other hand, in recent years, many nations, such as the United States in 1977, Japan in 1985, the United Kingdom in 1986, New Zealand in 1990, South Korea in 1991, Canada in 1993, and the countries of the European Union in 1998, have introduced competition.

Economists have been a consistent part of the debate, and a large number of empirical papers utilising econometric cost functions have led, followed and often criticised competition. The influential initial findings of Fuss and Waverman (1981a, 1981b), Fuss (1983), and Evans and Heckman (1983, 1984, 1986, 1988) were that competition in toll services was feasible. A number of other authors have concluded that competitive toll provision is inefficient (Charnes, Cooper and Sueyoshi (1988), Gentzoglanis and Cairns (1989), Pulley and Braunstein (1990), Röller (1990a,b), Ngo (1990), Diewert and Wales (1991a,b). Other authors have, more recently, used a variety of techniques to examine the costs of the provision of local services (Shin and Ying, 1992; Gabel and Kennet, 1994; Majumdar and Chang, 1998). These authors have reached conflicting conclusions concerning the desirability of competition in the provision of local services. Shin and Ying (1992) and Majumdar and Chang (1998) broadly support local competition, while Gabel and Kennett (1994) and Wilson and Zhou (2001) take the opposite view. In this survey we evaluate this conflicting evidence at both the toll and local service levels, and investigate the extent to which it is relevant to the kinds of changes in the market structure of the telecommunications industry which have been made.

1.2. Regulation and price-caps: The need for assessing productivity

Cost functions are also used for other important issues in telecommunications. They have proved important in establishing the *form* of regulation. First they were used to show the inefficiency of rate-of-return regulation. Now they are used to determine the crucial x factor. This is the rate of productivity improvement in price-caps proceedings[3]. Economies of scale and scope, allocative efficiency, and technical advance (productivity improvement) are all parts of the same puzzle. Therefore, determining efficiency or relative efficiency and the 'x' factor requires disentangling all the impacts of cost changes – scale, scope, innovation, regulation and factor price changes.

It is this disentanglement that is inherently difficult. The developers of the growth accounting literature (Denison, 1985; Solow, 1957; Jorgenson and Griliches, 1967), to name but a few, did not consider the problems associated with

[3] The conceptual and empirical relating to price-cap regulation are covered by Sappington in the chapter of 'Price Regulation' in this *Handbook*, and we do not go into details of this topic.

applying their methodology to an industry characterized by a high proportion of fixed and common costs, rapid technical change, and overt and changing government regulation.

A simple example indicates the nature of the problem:
Total factor productivity growth (*TFP*) is traditionally measured as the difference between Divisia indices of aggregate measures output and input growth (Diewert, 1976). *TFP* is the ratio of aggregate output Y to aggregate input X, so that:

$$\overset{o}{TFP} = \overset{o}{Y} - \overset{o}{X} \tag{1}$$

and

$$\overset{o}{Y} = \sum_j \frac{P_j Y_j}{R} \cdot \overset{o}{Y_j} \tag{2}$$

where P_j = of price of output j
Y_j = quantity of output j
$\overset{o}{Y}j$ = proportionate rate of growth of output j

$$R = \sum_j P_j Y_j = \text{total revenue}$$

and

$$\overset{o}{X} = \sum_i \frac{w_i X_i}{C} \cdot \overset{o}{X_i} \tag{3}$$

where w_i = price of input i
X_i = quantity of input i
$\overset{o}{X_i}$ = proportionate rate of growth of input i, and

$$C = \sum_i w_i X_i = \text{total cost.}$$

The aggregate production function takes the form

$$F(Y_1, \ldots, Y_m; X_1 .. X_n; t) = 0 \tag{4}$$

where F is the production function and t is a technical advance indicator, or

$$C = G(w_1 \ldots\ldots\ldots\ldots w_n;\ Y_1, \ldots\ldots\ldots\ldots Y_m; t) \tag{5}$$

where $G(..)$ is the cost function.

Totally differentiating (5) and re-arranging yields,

$$-\overset{o}{B} = \sum_j \varepsilon_{CY_j} \overset{o}{Y}_j - \sum_i \frac{w_i X_i}{C} \cdot \overset{o}{X}_i \tag{6}$$

where
$-\overset{o}{B}$ = technical change (the proportionate shift in the cost function with time), and ε_{CY_j} = cost elasticity of the jth output.

If production is characterised by constant returns to scale and all individual outputs are priced in the same proportion relative to marginal cost, $\varepsilon_{CY_j} = p_j Y_j / R$ and $\Sigma \varepsilon_{CY_j} = 1$.

$$\text{Then}: -\overset{o}{B} = \overset{o}{Y} - \overset{o}{X} \tag{7}$$

$$\text{Or} -\overset{o}{B} = T\overset{o}{F}P. \tag{7a}$$

However, with increasing returns to scale (decreasing returns to costs), Denny, Fuss and Waverman (1981) have demonstrated that measured *TFP* growth (or the 'x' factor) consists both of true technical change *plus* returns to increased scale. The relevant formula is

$$T\overset{o}{F}P = -\overset{o}{B} + (1 - \Sigma \varepsilon_{CY_j})\overset{o}{Y}. \tag{7b}$$

Therefore, to estimate *TFP* growth requires estimates of ε_{CY_j} and hence knowledge of the cost function. We return to the issues of measuring technical change and the 'x' factor in Section (3) below.

One issue is very important to raise here in this introduction. In the parametric analysis of the production structure, the choice of the functional form of the cost function has proven to be crucial. The majority of the analyses utilize a variant of the transcendental logarithmic (translog) cost function. The translog function is a second order approximation to an unknown cost function that imposes a minimal number of prior constraints on the specification of the underlying technology. This fact, and the additional fact that the estimating equations that result from applying the translog functional form are linear in parameters, has made it a popular function. However, a number of studies (Fuss, 1983; Röller, 1992) identify problems with the translog function as a flexible approximation. These problems can lead to biased estimates of economies of scale and scope (and of the 'x' productivity factor) if the true underlying cost function is poorly approximated by the translog function over the range of data that is of interest.

One possible solution is to estimate a functional form that contains the translog function as a special case (for example, the Box-Cox function [see Fuss, 1983]). But this does not avoid the functional form issue. Another solution is to search for

methods that do not rely on a chosen functional form as being the correct representation of the true unknown technology. A functional form is not imposed on the data. Rather, the data are left to 'speak' for themselves. Hence, the use of procedures which are non-parametric[4]. These techniques however raise other problems – namely the robustness of the estimation and their underlying statistical properties.

1.3. Efficiency measurement and the use of alternative approaches

A third general area of analysis has been the examination of the efficiency of the operations of specific firms. This is feasible with the use of goal programming and the allied Data Envelopment Analysis (DEA) techniques[5]. These belong to a family of non-parametric estimation methods useful for evaluating production conditions. Our emphasis in this chapter is on econometric techniques, *per-se*. Thus, we only briefly describe DEA and other non-parametric techniques[6].

As far as we are aware, Charnes, Cooper and Sueyoshi (1986) were the first to utilize non-parametric techniques in the estimation of cost functions for the telecommunications sector. Rather than using some form of average curve fitting of a specific functional form, using ordinary least squares (OLS) or maximum likelihood estimation (MLE) techniques, they utilized goal-programming/constrained regression techniques to examine the nature of subadditivity in the Bell system.

Since then, the use of DEA for efficiency analysis has been done with primarily two purposes in mind. First, to determine whether regulated firms were on the least cost expansion path (Majumdar, 1995). Second, to benchmark various dimensions of performance and production characteristics (Majumdar and Chang, 1996). DEA is a procedure that fits a curve to the outside (convex hull) of the data points. It is one of a class of techniques that define the maximal outputs for given levels of inputs[7]. The firm (s) at the outside envelope are defined as efficient; those with lower output for identical levels of inputs are relatively "inefficient."

[4] Non-parametric techniques such as DEA (Data Envelopment Analysis) make no assumption as about functional form other than convexity and piece-wise linearity. This property is extremely useful since functional relations between various inputs and outputs can be very difficult to define, if the assumptions of causal ambiguity and heterogeneity at the firm level (Lippman and Rumelt, 1982) are made.

[5] See Banker, Charnes and Cooper (1984) and Charnes, Cooper and Rhodes (1978) for details of the primary algorithms.

[6] See Seiford (1996) and Seiford abd Thrall (1990) for more reviews for the DEA literature.

[7] Another technique in this class is the frontier cost or frontier production function technique. An important difference between DEA analysis and frontier cost function estimation is that the former is a non-parametric technique, whereas the latter is a method of estimating the parameters of a specific function. However, DEA can be used with multiple outputs, while the use of the frontier cost or production functions is restricted to the single output case. See Jha and Majumdar (1999) for an application of parametric frontier cost estimation techniques in the analysis of telecommunications sector issues.

Following the literature:

Identify u_r, v_i as weights on output y_r and input x_i, which we calculate to maximize the efficiency rating h_k of the kth firm[8].

$$h_k(u, v) = \frac{\sum_{r=1}^{s} u_r y_{rk}}{\sum_{i=1}^{m} v_i x_{ik}}. \tag{8}$$

The derived linear mathematical programming problem is:

$$\max_{\mu, v} w_k = \sum_{r=1}^{s} \mu_r y_{rk} \tag{9}$$

subject to

$$\sum_{i=1}^{m} v_i x_{ik} = 1$$

$$\sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leqslant 0 \quad (j = 1, 2, \ldots\ldots \mathrm{k}, \ldots\ldots \mathrm{n}) \tag{10}$$

$$\mu_r \geqslant 0 \quad (r = 1, 2, \ldots\ldots \mathrm{s})$$

$$v_i \geqslant 0 \quad (i = 1, 2, \ldots\ldots \mathrm{m})$$

where

$$\mu = \frac{u}{v} x_k; \quad v = \frac{v}{{}_v x_k} \tag{11}$$

and
w_k is the 'efficiency' score of the kth firm.

Note that each observation, be it for firm m relative to firm c, or for firm c, or for period t relative to period $t-1$, a parameter is generated. In its earlier uses, statistical properties of these parameters were not well identified. Thus, the robustness of the results is a clear issue. Likewise, the choice of weights μ, v are arbitrary in the sense that these are empirically generated by the estimation process[9].

[8] Other constraints are incorporated so that each efficiency ratio < 1.

[9] All these techniques minimize something other than the least squares of the residuals. Minimizing the sum of the absolute errors (MSAE) or commonly LAR (least absolute residual) goes back to Edgeworth (Maddala, 1977). Charnes, Cooper and Ferguson (1955) showed that minimizing LAR is equivalent to this linear maximization problem denoted in (8) and (9) above.

There are two important issues involved in the use of DEA techniques to analyze the efficiency characteristics of production units. First, the results can be very sensitive to outlier observations, which may be the result of measurement error in the data. Some progress has been made in identifying potentially influential outliers (see Wilson, 1993) which could bias the estimation of efficiency scores. The second main issue associated with the majority of applications of DEA is the lack of a statistical framework from which to estimate the finite sample bias and variance of the efficiency scores. Recently, however, Simar and Wilson (2000) have derived a method of estimating the statistical properties of the efficiency measures for a multiple output, multiple input production process using bootstrapping techniques. To our knowledge, these newer advances have not been applied as yet to the study of efficiency in telecommunications.

1.4. General comments

Most econometric studies examine historical data for the telecommunications industry in the U.S. or Canada, and for specific carriers such as AT&T and Bell Canada. Correspondingly, our review is focussed primarily on these two contexts. The studies examine the actual range of services offered by the industry and by these carriers. There are several reasons to conclude that cost function analyses were not robust for the initial public policy purposes of the 1970's and 1980's, which was to determine the socially correct degree of entry into each service or across services or even the 'divestiture' of AT&T itself. First, the US and Canadian industries were organised as regional monopolies until recently. AT&T provided about 80 percent of the U.S. telecommunications services until 1984. Bell Canada had a legal monopoly over most of Quebec and Ontario provinces until 1992. Hence, the data that the econometrics relied on were for a firm that provided a set of services and existed as a regulated monopoly with strict entry barriers.

Second, there were few carriers who provided single services. The one empirical result consistent across most studies, including early ones, is the existence of economies of scale in the provision of telecommunications services *in the aggregate*. This provides no indication as to whether or not individual services such as toll or local service should be monopolised or not. Third, the econometric results that conceptually could provide such evidence, those pertaining to toll-specific economies of scale and existence of natural monopoly (established by the subadditivity of the cost function), do not withstand careful scrutiny.

In order to specify tests of service-specific economies of scale and subadditivity relevant to the question at hand, the researcher must be able to estimate the technology of a firm that produces only that service. Little actual data exist in any country, with the exception of very recent post-entry data, which pertain to such a 'stand-alone' firm. In any case, such data have rarely been used in econometric estimations of the telecommunications production technology.

This lack of appropriate data has forced researchers into one of two straight-jackets. The first has been to concentrate on analyses that are not relevant to the emerging market structures. These structure are generally characterised by a multi-product firm facing a firm, or firms, which provide only toll service and are interconnected with the multi-product firm which provides both local as well as toll services. The second has led to an incorrect specification of the cost function, which represented the stand-alone technology, and to the projection of the estimated function far outside the actual experiences of the firms whose data sets were used for estimation.

2. The concepts of economies of scale, scope, and subadditivity

This section describes the econometric approaches to evaluating cost and production functions. We discuss the concepts of economies of scale, economies of scope, and subadditivity. These are concepts central to an understanding of the econometrics literature.

2.1. Economies of scale and scope

Before proceeding to a more detailed discussion of the actual econometric studies, it is useful to provide a brief outline of the concepts of economies of scale, economies of scope, and subadditivity, concentrating on the applicability of these concepts to the question of the desirability of competition in the provision of a single telecommunications service. Due to the intense interest in this and other public policy questions that require knowledge of the telecommunications production technology, several extensive reviews of the econometrics literature have been published (see especially Fuss, 1983; Kiss and Lefebvre, 1987; and Waverman, 1989). Each of these studies provides a detailed description of the concepts of scale, scope and subadditivity as they have been used in the econometrics literature. Here to simplify the discussion we concentrate on their applicability to the first main question addressed – the competitive provision of toll services.

These concepts can most easily be understood in terms of a firm's cost function. This is the function that is almost always estimated in contemporary analyses of the telecommunications production technology. The cost function of a telecommunications firm is defined as the firm's minimum cost of providing a set of telecommunications services, conditional on given purchase prices of the factors of production and on a particular state of technology (see Equation 5). In this section we simplify the mathematical presentation by deleting these conditioning variables from explicit consideration. Also for simplicity we will restrict ourselves to the case where only two outputs are produced: toll service (Y_T) and local service (Y_L). For a firm that produces both toll and local services, the cost function takes the form

$$C = C(Y_T, Y_L). \tag{12}$$

For a firm which produces only toll services,

$$C^T = C^T(Y_T) \tag{13}$$

and for a firm which produces only local services,

$$C^L = C^L(Y_L) \tag{14}$$

Note that by superscripting C in Equations (13) and (14) we have specified, as seems reasonable, that firms which produce only toll services or only local services have production technologies that differ from the production technology of a firm which produces toll and local services jointly. This difference is sufficiently fundamental that these technologies cannot be recovered from Equation (12) simply by setting the missing output equal to zero. This observation is very important since it means that (13) or (14) can never be recovered from data generated by a firm represented by (12). Unfortunately, the assumption that (13) or (14) can be recovered from (12) is maintained in all the econometrics literature which considers this issue explicitly.

When will the specifications of (12), (13) and (14) be appropriate, and the specification used in the econometrics literature incorrect? One situation in which this will occur for most forms of the cost function is if there are fixed costs in the provision of services, but these fixed costs differ depending on whether the firm supplies toll, or local, or both. Another situation arises if the value of marginal cost depends on whether the firm corresponds to (12), (13), or (14).

One can illustrate the preceding argument by specifying three very simple cost functions for C, C^T, and C^L. Suppose the joint production technology, toll-only technology, and local-only technology cost functions take the forms[10]:

$$C(Y_T, Y_L) = F_{TL} + \alpha_T Y_T + \alpha_L Y_L + \alpha_{TL} Y_T Y_L \tag{15}$$

$$C^T(Y_T) = F_T + \beta_T Y_T \tag{16}$$

$$C^L(Y_L) = F_L + \gamma_L Y_L \tag{17}$$

where F_{TL}, F_T and F_L are the fixed costs associated with the three types of technologies respectively, and α, β and γ are cost function parameters. If we used

[10] We ignore technical change in this section for simplicity.

Equation (15) to specify the toll-only and local-only cost functions, we would obtain

$$C(Y_T, 0) = F_{TL} + \alpha_T Y_T \tag{18}$$

$$C(0, Y_L) = F_{TL} + \alpha_L Y_L \tag{19}$$

Equations (18) and Equation (16) will not be identical, as required, since the fixed costs of the joint production technology will not normally equal the fixed costs of the toll-only technology ($F_{TL} \neq F_T$), and the marginal costs will also not be equal ($\alpha_T \neq \beta_T$). A similar argument holds for the case of the local-only technology, since $F_{TL} \neq F_L$ and $\alpha_L \neq \gamma_L$.

2.1.1. Overall economies of scale

A multi-product firm is said to be subject to overall increasing (decreasing) returns to scale if a proportionate increase in all outputs results in a less (more) than proportionate increase in total costs. Since the concept of overall returns to scale for a multiple output technology is defined only when all outputs change in proportion, this concept is not directly relevant for evaluating the cost implications of the situation that was considered in much of the literature, where a telephone company would share the toll market with competitors but remain a monopolist in the local market. As we will show below, the relevant concept for this case is a version of subadditivity – the basic technical property of a cost function that signals the presence of natural monopoly characteristics. It is also well known that the existence of overall economies of scale by itself is neither necessary nor sufficient for subadditivity (see, for example, Panzar, 1989, page 25)[11].

2.1.2. Economies of scope

The second concept of interest is that of economies of scope. A firm which produces both toll and local services is said to enjoy economies of scope if it can produce these services at lower cost than would occur if each service were

[11] Despite this theoretical result, the fact that the majority of econometric studies find increasing overall returns to scale is often cited as evidence in favour of natural monopoly. This is an incorrect inference. The evidence on overall returns to scale does suggest that increasing returns and/or economies of scope may exist somewhere in the network. Most observers trace the source to the physical capital requirements of providing subscriber access to the network. Since telephone companies' accounting procedures typically allocate access outputs to the local service aggregate, the local service output is the likely location of any cost-reducing advantages of larger size found in the econometric estimates. Access is also the probable source of any cost complementarities between local and toll service aggregates.

produced separately by a stand-alone firm. We can measure economies of scope by the variable *SCOPE*, where

$$SCOPE = [C^T(Y_T) + C^L(Y_L) - C(Y_T, Y_L)]/C(Y_T, Y_L). \tag{20}$$

If we multiply *SCOPE* by 100, it measures the percentage change in total cost of stand-alone productions of toll and local services compared with the cost of joint production of the two services by a single firm. If $SCOPE > 0$ economies of scope exist, whereas if $SCOPE < 0$ diseconomies of scope exist.

The above ideas can be illustrated using our simple cost functions (15), (16) and (17). If we substitute these functions into Equation (20), the numerator of (20) becomes:

$$NUM = (F_T + F_L - F_{TL}) + (\alpha_T - \beta_T)Y_T + (\alpha_L - \gamma_L)Y_L - \alpha_{TL}Y_TY_L \tag{21}.$$

SCOPE is greater than zero when *NUM* is greater than zero. *NUM* will more likely be greater than zero when joint production yields savings in total fixed costs ($F_{TL} < F_T + F_L$), and when joint production yields cost complementarities ($\alpha_{TL} < 0$). The signs of the other terms in (21) depend on the marginal cost relationships among the joint and stand-alone technologies, and could plausibly be either positive or negative.

Although Equation (21) corresponds to the definition of *SCOPE* given by Equation (20) for our simple cost functions, it does not correspond to what would be calculated in the econometrics literature. First, all studies in the econometrics literature which attempt to measure economies of scope assume that $F_{TL} = F_T = F_L$, so that *all* the fixed costs associated with the provision of both toll and local services would be borne by each stand-alone firm. This clearly overstates the fixed cost contribution to a positive *SCOPE*, since it assumes that having separate toll and local service-providing firms will necessarily lead to complete duplication of facilities. It is more likely that some facilities will be jointly used. For example, when Mercury Communications entered the toll market in the United Kingdom, it did not construct its own local loops but shared British Telecom's local service facilities through interconnection. Second, the econometric studies normally implicitly assume that the second and third terms in (14) are identically equal to zero, since there would be no provision for the possibility of distinct stand-alone technologies.

There is a third issue that is illustrated by our simple cost functions. The presence of the cost complementarity term $\alpha_{TL}Y_TY_L$ in *NUM* implies that estimates of *SCOPE* using (13) will be calculated under the assumption that the *entirety* of any cost complementarity that may exist would be lost if the provisions of toll and local services were separated. This assumption is another source of overestimation of possible lost economies of scope, since interconnection of facilities would substantially mitigate this effect.

The above discussion illustrates the importance of interconnection provisions and the proper pricing of these provisions, when competition with an incumbent telecommunications operator is permitted, in order that any network economies of scope that may be present not be lost.

2.1.3. Product-specific returns to scale

The third concept that is of interest is the concept of product-specific returns to scale. Suppose we measure the incremental cost of producing toll services for a firm that already produces local services as:

$$IC_T = C(Y_T, Y_L) - C^L(Y_L). \tag{22}$$

Then toll-specific returns to scale is measured by

$$S_T = IC_T/(Y_T.MC_T) \tag{23}$$

where MC_T is the marginal cost of supplying toll services when the firm produces both toll and local services. If $S_T > 1$ the firm is said to produce toll services with increasing returns to scale, if $S_T < 1$ toll-specific returns to scale are said to be decreasing, and if $S_T = 1$ toll-specific returns to scale are said to be constant.

Similarly, local-specific returns to scale is measured by

$$S_L = [C(Y_T, Y_L) - C^T(Y_T)]/(Y_L.MC_L) \tag{24}$$

There is an interesting and illuminating relationship between overall returns to scale, product-specific returns to scale, and economies of scope given by the equation (see for example Panzar [1989]):

$$S = [a_T.S_T + a_L.S_L]/(1 - SCOPE) \tag{25}$$

where $a_T = (Y_T.MC_T)/(Y_T.MC_T + Y_L.MC_L)$ and $a_L = (Y_L.MC_L)/(Y_T.MC_T + Y_L.MC_L)$ are cost elasticity shares for toll and local services respectively, and S is the measure of overall returns to scale.

From Equation (25) we can see why a finding of increasing overall returns to scale tells us nothing of relevance for the question of whether competition should be allowed in the provision of toll services. There are numerous combinations of product- specific returns to scale and returns to scope characteristics that could result in increasing overall returns to scale[12].

[12] One situation which would unambiguously favour competition in the provision of toll services is the scenario where $S_T < 1$ (toll-specific diseconomies of scale), $SCOPE = 0$ (absence of economies of scope) and $S_L > 1$ (local-specific economies of scale). If local- specific economies of scale are large and toll-specific diseconomies of scale are small, this configuration would likely produce a finding of overall economies of scale ($S > 1$).

2.1.4. Subadditivity

The fourth concept of interest is that of subadditivity. This is the concept most closely connected to the question of whether a firm is a natural monopolist in the provision of the services demanded in the marketplace. Hence, it is most directly relevant to the issue of competitive provision of toll services. A telephone company will be a natural monopolist in the provision of toll and local services if it can provide the levels of services demanded at lower cost than can be provided if alternative sources of supplies are permitted. This must be true for all possible levels of services. In this case, we say that the cost function of the monopolist is subadditive. For the situation of one versus two firms found in the econometrics literature, we can measure the degree of subadditivity from the equation (see Evans and Heckman [1983]):

$$SUB = \{C(Y_T, Y_L) - [C(\phi.Y_T, \omega.Y_L) + C((1-\phi).Y_T, (1-\omega).Y_L)]\}/C(Y_T, Y_L). \tag{26}$$

When SUB is multiplied by 100, it measures the percentage difference in total industry costs between single firm production of toll and local services and any two firm configuration. The particular two firm configuration is determined by the weights ϕ and ω which designate the way in which industry output is divided between the two firms. If $SUB < 0$ for *all possible values of* ϕ *and* ω the industry cost function is subadditive and single firm production is cost-minimizing. The condition is stringent. If *any* values of ϕ and ω can be found for which $SUB > 0$, monopoly provision of both toll and local services is *not* cost-minimizing.

The test of subadditivity based on Equation (26) is a global test, and is both necessary and sufficient for the existence of a natural monopoly. However, this test requires that the cost function be valid for values of outputs well outside the range of actual experience, including stand-alone production of both toll and local services (e.g. when $\phi = 1$ and $\omega = 0$ or $\phi = 0$ and $\omega = 1$). It also requires that when only one output is produced by a firm, $C^T(Y_T) = C(Y_T, 0)$ and $C^L(Y_L) = C(0, Y_L)$. This, as noted previously, is unlikely to be correct.

To counteract the criticism associated with the global test, i.e., that the cost function must be valid for outputs outside the range of actual experience, Evans and Heckman (EH) developed a local test of subadditivity, in which they restricted ϕ and ω to values such that neither of the two firms in the competitive scenario would produce outside of the historically-observed range of the monopolist. The EH test has been used in all subsequent tests of natural monopoly (except Röller [1990a] and Pulley and Braunstein [1990]). However the EH test is only a necessary test, and because it is *not*

sufficient, can *only* be used to reject the existence of natural monopoly, not confirm it[13].

Aside from the limited nature of the evidence that the EH test generally provides, it has an important flaw in the present circumstances. The data region in which one firm provides both toll and local services, and other firms provide only toll services is outside the data region used to estimate telecommunications cost functions. Thus the EH test is undefined for the case of competitive entry normally under consideration and results based on this test were not relevant to the kinds of competitive entry taking place around the world in the 1980's and 1990's[14].

A variant of the subadditivity test *is* relevant for the type of competitive entry that is being analyzed. Suppose we replace Equation (26) with an alternative definition of SUB, denoted SUB_T:

$$SUB_T = \{C(Y_T, Y_L) - [C^T(\phi.Y_T) + C((1-\phi).Y_T, \phi.Y_T, Y_L)]\}/C(Y_T, Y_L). \tag{27}$$

Comparing (27) with (26) we see that there are three differences. First, in (27) $\omega = 0$ (one firm produces no local output). Second, the cost function for producing toll output alone has a superscript T to indicate, as noted previously, the different nature of the toll stand-alone technology. Third, the toll output of the stand-alone firm enters the cost function of the other firm as an economies of scope cost-reducing externality, since, as also noted previously, the two firms will be interconnected and hence share facilities such as terminal equipment and local loops. This sharing of facilities will mitigate cost increases that are associated with a loss of economies of scope. These will be suffered by the joint output-producing firm via the diversion of toll output to the other firm. Of course the cost function for the toll-only producing firm will have to include any interconnection costs which are caused specifically by the entry of the competitor.

If $SUB_T < 0$ for all values of ϕ such that $0 < \phi < 1$, then single firm production is cost-minimizing, otherwise competitive provision of toll services does not raise supply costs. This is the correct subadditivity (or natural monopoly) test for the

[13] In the words of Evans and Heckman (1983, p. 36): "Our test is one of necessary conditions for natural monopoly. Rejection of that hypothesis is informative. Acceptance within a region ... is not informative. All that acceptance demonstrates is that there are inefficient ways to break up AT&T. Failure to reject the necessary conditions for natural monopoly does not imply that AT&T was a natural monopoly ....To perform such a test requires extrapolation of the cost function well outside the range of the data used to estimate it."

[14] The EH test was developed in conjunction with the anti-trust litigation which led to the divestiture of AT&T. It was designed to analyze a market structure in which competing toll-producing firms would interconnect with local-producing firms that would not be in the inter-state toll provision business. This kind of market structure is unique to the United States. In other countries, the market structure analyzed in this paper, where toll-producing firms compete with a firm (the former monopolist) that produces both toll and local services and provides facilities for interconnection, is the relevant market structure.

version of competitive entry that is relevant outside of the U.S. (see footnote 8). Unfortunately, SUB_T has never been calculated. This is perhaps not surprising since the calculation requires knowledge of the toll stand-alone technology and the economies of scope effect of interconnection. Only Röller (1990a) performs a global test as is required. However, the use of his results to compute SUB_T would bias the test towards a finding of subadditivity for two reasons. First, he assumes the amount of fixed costs would be the same for a toll stand-alone technology as they are estimated to be for the monopoly firm supplying both toll and local services. Second, Röller's model implies all economies of scope associated with the toll-only firm's output would be lost if two firms operated in the toll market while one firm supplied all the local output, contrary to the reality of interconnected facilities.

The calculation of SUB_T would provide the correct *static* test of the natural monopoly hypothesis, but even this calculation would fail to take into account the dynamic effects of competitive entry on the costs of service provision. In particular, there are three effects, which imply that the correct static test is itself biased in favour of a finding of natural monopoly when the dynamic setting is ignored. First, the static test assumes the level of total demand is the same pre- and post-competition. To the extent that competition-induced market stimulation leads to expanded markets, and there exist economies of scale and scope, both the incumbent and entrant's unit costs will be overstated by the static test. Second, competition will likely act as a spur to increased efficiency on the part of the incumbent, or a reduction in x-inefficiency (Majumdar, 1995), and for this reason the incumbent's post-entry costs are overstated by the static test. Finally, the static test assumes the same state of technology pre- and post-competition. To the extent that technical change in the industry is accelerated by the presence of competition, the static test overstates the post-competition costs of both the incumbent and entrant firms.

On the other side of the ledger, the static test does not consider any inter-temporal costs associated with competition, such as the potential ongoing interconnection costs and increased marketing costs that are specific to the existence of a competitor.

3. Technological change[15]

As noted in the introduction, the analysis of technological change has been crucial in telecommunications as cost reductions come about through changes in factor prices or technical change. Thus, looking at technical change issues in a formal way is useful. Technical advances have been a hallmark of the telecommunications sector, since the sector relies on both switching equipment, using computers and semi-conductors, and transmission equipment using radio and multiplexing.

[15] This section draws heavily from Fuss (1994).

Technological change has affected various services differently. Hence, the possibility of efficient entry in separate service markets was differentialy affected by technical progress. With the replacement of rate-of-return regulation with price-caps regulation, the measurement of technical progress (the expected productivity enhancement) is a crucial aspect of the regulatory regime, and econometric measurement of technical progress is a part of the process.

The conventional Törnqvist index[16] for measuring *TFP* growth between years $t-1$ and t is calculated from the log difference formula:

$$\Delta \log TFP^R = \Delta \log Y^R - \Delta \log X \tag{28}$$

where

$$\Delta \log Y^R = \Sigma(1/2)(R_{jt} + R_{j,t-1}).[\log Y_{jt} - \log Y_{j,t-1}] \tag{29}$$

$$\Delta \log X = \Sigma(1/2)(S_{it} + S_{i,t-1}).[\log X_{it} - \log X_{i,t-1}] \tag{30}$$

where

Y_{jt} is the amount of the *ith* output produced at time t,
X_{it} is the amount of the *ith* input utilized at time t,
R_{jt} is the revenue share of the *ith* output in total revenue, and
S_{it} is the cost share of the *ith* input in total cost.

The superscript R indicates that revenue weights are used in the calculations. $\Delta \log Y^R$ and $\Delta \log X$ are often referred to as the rates of change of aggregate output and aggregate input respectively, since they result from procedures which aggregate the rates of change of individual outputs and inputs.

It has been recognized since the late 1970s that a crucial assumption used in establishing the linkage between $\Delta \log TFP^R$ and the annual change in production efficiency is that output prices be in the *same* proportion to marginal costs for all outputs at any point in time. This is an inappropriate assumption in the case of telecommunications firms. For these firms the price of toll output, as a broad service category, has exceeded the marginal cost of toll production, and the local service price, including the access price[17], is less than the relevant marginal cost. Empirical support for these assertions can be found in Fuss and Waverman (1981a, 1981b), where Bell Canada prices are compared with econometric estimates of marginal costs. This pattern of cross-subsidization eliminates any possibility that the proportionality assumption could be satisfied in the historical data.

[16] The Törnqvist index is a discrete time empirical approximation to the continuous Divisia index introduced earlier.

[17] Access is included as one of the outputs in the sub-aggregate 'local' in the productivity accounts of most telephone companies.

Fuss (1994) formally demonstrated the fact that Equation (28) only measures production efficiency growth when the price/marginal cost proportionality assumption holds. He also demonstrated that, when this assumption does not hold, the correct form of the *TFP* index for a cost-minimizing firm is

$$\Delta \log TFP^C = \Delta \log Y^C - \Delta \log X \tag{31}$$

where

$$\Delta \log Y^C = \sum \frac{1}{2}(M_{jt} + M_{j,t-1}).[\log Y_{jt} - \log Y_{j,t-1}] \tag{32}$$

and M_{jt} is the cost elasticity of the *jth* output divided by the sum of the cost elasticities, summed over all outputs ($M_{jt} = \varepsilon_{CY_j}/\Sigma\varepsilon_{CY_j}$). M_{jt} is denoted a 'cost elasticity share' to distinguish it from the cost elasticity itself.

Whether the average cost elasticities[18] or average cost elasticity shares are the correct weights for weighting the rates of growth of the individual outputs depends on the definition of *TFP* utilized. If productivity growth associated with scale economies is excluded from the definition, the correct weight is the cost elasticity. This will be the case when *TFP* growth and technical change are by definition synonymous. If scale economies are included as a potential source of *TFP* growth, the correct weight is the cost elasticity share[19]. Both definitions of *TFP* growth can be found in the productivity literature.

Which cost elasticity shares to use depends on whether one believes a short-run model, with capital quasi-fixed, or a long run model, with capital variable, is appropriate. From the point of view of *TFP* growth measurement, the difference in model implies a difference in the valuation of the capital input, both in the calculation of the output cost elasticity aggregation weights and the input aggregation weights. For the long-run model, the capital input is valued at its user cost. For the short-run model, the price of capital services to be used is its shadow value at the point of temporary equilibrium[20].

[18] See Caves and Christensen (1980).

[19] When production is subject to constant returns to scale, scale economies do not contribute to *TFP* growth and the two possible definitions of *TFP* growth coincide. Also, this is the case where the cost elasticity share and the cost elasticity are identical, since with constant returns to scale the cost elasticities sum to unity. The above statements are directly applicable to a long run equilibrium analysis. With respect to a short run equilibrium model, these statements remain valid as long as the output cost elasticities are based on shadow price valuation of the quasi-fixed inputs.

[20] Most *TFP* growth calculations for telecommunications are based implicitly on a long-run model, since the user cost of capital services is used for the capital input price. But, as Bernstein (1988, 1989) has emphasized, given the capital-intensive nature of the production process, it may be that the short-run model is more appropriate. In the single output case, Berndt and Fuss (1986) have shown that the short-run model can be implemented through the replacement of the user cost of capital by a shadow value in the calculation of the input aggregation weights. Berndt and Fuss (1989) extend the analysis to the multi-output case, and demonstrate that the appropriate procedure is, in addition, to replace the user cost with the shadow price in the calculation of the cost elasticity shares.

3.1. Estimating marginal costs

There are two procedures that have been used to estimate marginal cost, or equivalently the cost elasticity share. The first procedure is to estimate an econometric multiple output cost function. This procedure has been used by Denny, Fuss and Waverman (1981), Caves and Christensen (1980), Caves, Christensen and Swanson (1980), and Kim and Weiss (1989), among others. The second procedure is to utilize the results of cost allocation studies to approximate the cost elasticity weights. This procedure has not been utilized previously in the telecommunications sector, but was used by Christensen et al. (1985, 1990) in their analysis of the United States Postal Service total factor productivity[21].

Estimation of a multiple output cost function would appear to be an obvious way to obtain the needed cost weight information. The logarithmic derivatives of the cost function with respect to the individual outputs provide the cost elasticities needed to construct the cost elasticity weights. However, this method is not without problems. For both the United States and Canada it has proven difficult to obtain well-behaved multiple output telecommunications cost function estimates with positive cost elasticities. This appears to be especially true when data for the 1980s are added to the sample[22]. Even when cost functions with satisfactory theoretical properties have been obtained, the economic characteristics estimated have remained a subject of controversy (see Kiss and Lefebvre, 1987; and Fuss, 1992). Finally, with econometric cost functions, at most three aggregate cost elasticities can be obtained, whereas, for example, for the cost allocation data used in Fuss (1994), seven cost elasticities can be obtained, permitting a more disaggregated analysis[23]. In a similar kind of setting, Christensen et al. (1985) proposed as a practical empirical approximation to the required marginal cost datum the average (unit) allocated cost obtained from cost allocation studies. The use of cost allocation studies relies on the fact that the methodology of cost allocation leads to careful attempts to allocate costs to service categories that are causally

[21] The cost allocation data has also been used by Curien (1991) to study the pattern of cross subsidies in the Canadian telecommunications industry. His paper contains an example of the kind of information which is available from typical cost allocation studies of Bell Canada and British Columbia Telephones (see Table 1 of his paper).

[22] The two papers of which we are aware that estimate cost functions using Canadian data from the 1980s (Gentzoglanis and Cairns [1989], Ngo [1990]) are plagued with lack of regularity and/or cost elasticity estimates that are negative. Highly trended output data and inadequate technical change indicators appear to be particularly problematic with respect to the 1980s data. For discussions of difficulties with the U.S. data see Waverman (1989), Röller (1990) and Diewert and Wales (1991a). This issue has not been investigated using the cross-section time series data for local exchange carriers contained in studies such as Shin and Ying (1992) and Krouse et al. (1999).

[23] In practice, Fuss only used three cost elasticities due to limitations in the productivity data that were available.

related to the production of those services. In Canadian telecommunications cost studies, unlike those undertaken in the United States, not all costs are allocated, since it is recognized that some costs, the 'common costs,' cannot be allocated on a conceptually sound basis. The procedures adopted by the Canadian Radio-Television and Telecommunications Commission (CRTC) use peak traffic in the allocation of usage sensitive costs and hence the costs allocated are more closely related to incremental costs than is the case in the United States, where state-federal political considerations have often been important in allocation procedures.

It is well known that the use of allocated costs to proxy marginal costs can be problematic. Accounting procedures and economic causality do not always mesh. In addition, incremental cost may not be constant over the range of output considered. But the approximation has several advantages. The major advantage is that, despite the limitations of the cost allocation exercise, unit allocated costs can be expected to much more closely satisfy the proportionality requirement than prices, given the very large cross subsidization from toll to local services at the centre of most counties public policy toward telecommunications. Hence, *TFP* growth rates constructed from allocated cost weights should provide a more accurate picture of efficiency changes than TFP growth rates constructed from revenue weights.

The cost elasticity shares M_{jt} can be expressed in terms of original costs as

$$M_{jt} = (Y_{jt}.MC_{jt})/(\Sigma Y_{jt}.MC_{jt}) \tag{33}$$

where MC_{jt} is the marginal cost of the *jth* output at time *t*. Replacing MC_{jt} in (33) with a constant of proportionality times the average allocated cost (service category *j*) yields the alternative expression for M_{jt},

$$M_{jt} = \frac{costs\ allocated\ to\ service\ category\ j}{total\ costs\ allocated\ (excluding\ Common)} \tag{34}$$

$$= \text{allocated cost share (service category } j). \tag{35}.$$

3.2. *Revenue weights versus cost weights in price caps formulas*

As can be seen from the above analysis, there are two possible weights (R_{jt} or M_{jt}) that can be used to aggregate outputs in the *TFP* formula. If the purpose is to measure the growth of efficiency in the telecommunications firm's operations, the cost elasticity weights M_{jt} are the correct choice. However, the choice is not so obvious if the purpose is to estimate an '*x*' factor for use in a Price-Caps formula.

The Divisia aggregate index of the growth of output prices, as a matter of algebra, takes the form:

$$\textit{Output Price Index Growth Rate} = \textit{Input Price Index Growth Rate} - TFP^R. \tag{36}$$

An implication of Equation (36) is that the 'x' factor would be the revenue-weighted *TFP* growth rate.

Denny, Fuss and Waverman (1981) and Fuss (1994) demonstrated that the relationship between TFP^R and TFP^C could be expressed as (up to a second order of smallness due to discreteness)

$$\Delta\log TFP^R = \Delta\log TFP^C + (\Delta\log Y^R - \Delta\log Y^C). \tag{37}$$

Equation (37) implies that an alternative expression for the output price index growth rate is

$$\textit{Output Price Index Growth Rate} = \textit{Input Price Index Growth Rate} - TFP^C - (\Delta\log Y^R - \Delta\log Y^C). \tag{38}$$

An implication of Equation (38) is that the 'x' factor would be the cost-weighted *TFP* growth rate, adjusted for the difference between the revenue-weighted and cost-weighted aggregate output growth rates expected during the time period that the telecommunications firm would be subject to the price-caps regime.

While, as noted above, it is clear that when revenue-weighted and cost-weighted *TFP* growth rates differ historically, the cost-weighted measure is a superior indicator of past efficiency growth; it is not as clear which measure should be used in determining the productivity offset in a price-caps formula. This is because the productivity offset in a price caps formula measures more than efficiency changes. It measures the ability of the firm to sustain output price declines, net of input price inflation, without declines in profit. So, for example, if intensified competition causes a decline in the price-marginal cost margin of a service with a positive margin, and the output of that service does not increase sufficiently to offset the margin loss, there will be a reduced ability on the part of the firm to sustain a price index decline, even when efficiency growth is unchanged. This is the reason why, when the output price index is expressed in terms of the cost-weighted *TFP* measure, an additional term, $(\Delta\log Y^R - \Delta\log Y^C)$, must be included in the price-caps formula.

The correct conceptual choice, Equation (36) or Equation (38), depends on a comparison of the price and marginal cost relationship in the historical period, from which the estimate of total factor productivity growth is drawn, with the relationship expected to prevail in the price-caps period.

An example drawn from the case of two Canadian telephone companies, Bell Canada and British Columbia Telephone, may clarify the issue. During the 1980s, rates for toll calls exceeded marginal costs and rates for local calls were less than marginal costs. Fuss (1994) demonstrated that this condition, along with the more

rapid growth of toll, caused the revenue-weighted *TFP* growth measure to overestimate substantially efficiency growth. However the revenue-weighted measure might still be the appropriate *TFP* offset for a price caps plan for these companies. This would occur if the pattern of price, marginal cost relationships were to be continued in the price caps period and there were no significant expected changes in relative growth rates of outputs.

On the other hand the price caps period could represent a period of transition to marginal cost-based pricing. The price and marginal cost relationships would be maintained, but relative output growth rates in the price caps period were expected to differ substantially from the historical period (due to heightened competition in the provisions of toll services). In that case the conceptually correct 'x' factor would be a variable offset which combined the cost-weighted *TFP* measure, from the historical data, with an adjustment term that took into account the changing revenue, cost-weight differentials and the changing relative output growth rates (Equation (38)).

While the situation described in the last paragraph is a realistic view of what occurred when telecommunications markets were opened to competition, to our knowledge, the conceptually correct Equation (38) has never been implemented in actual practice. There appear to be several disadvantages from a policy perspective that need to be taken into account. First, cost elasticity weights would have to be calculated. In this chapter we have reviewed two methods for calculating such weights: econometric cost functions and cost separations data. In an actual regulatory hearing setting, the calculation would likely be controversial.

Second, a price caps formula based on Equation (38) would depend on the growth rates of outputs, which creates incentive problems. A telecommunications firm would be aware that a lower rate of growth of output for a service that provides a positive margin, or a higher rate of growth for a negative margin service, would result in a lower productivity offset.

It is interesting to note that had regulatory jurisdictions implemented a cost-weighted 'x' factor, that implementation would probably have been to the advantage of the telecommunications firms, in that it would likely have resulted in a lower 'x' factor as competition intensified. This would have occurred because competitors first targeted the firm's high margin services. This targeting resulted in reductions in the firm's price-marginal cost margins and a reduction in the output growth rates for these high margin services. Both impacts would mean that the term $(\Delta \log Y^R - \Delta \log Y^C)$ in Equation (38) would have declined, or possibly become negative, and the resulting offset would have be lower than was the case when a revenue-weighted output growth rate index was used.

4. A selective review of the evidence

Fuss (1983), Kiss and Lefebvre (1987), and Waverman (1989) have surveyed the econometric results on scale, scope and subadditivity through 1986. Appendix 1 is

Table 1
Estimates of returns to scale and scope[1]

Study	Overall economics of scale (s)	Toll-specific economies of scale (s_T)	Local-specific economies of scale (s_L)	Economies of scope (SCOPE)
Kiss *et al.*l (1981, 1983)[2]	1.62	1.09[3]	1.56	0.09[4]
Ngo (1990)[5]				
Model 1	2.10	1.02[6]	−1.19	1.44
Model 2	2.03	1.09[6]	−0.98	1.42
Model 3	1.58	0.33[6]	−4.19	3.56

Notes: 1. Estimates are for 1967, the year in which the data are normalized. 2. Results are for Kiss et al.'s preferred Generalized Translog model as reported in Kiss and Lefebvre (1987). 3. Insignificantly different from unity. 4. Insignificantly different from zero. 5. Estimates for product-specific economies of scale and economies of scope have been calculated by the present author. 6. Significance unknown.

an update of Table 8 of Waverman (1989) to include similar studies carried out through 1999. Only studies, which disaggregate output, are included. The one result that seems consistent is a finding of increasing overall returns to scale though Fuss and Waverman (1977, 1981a,b) and Elixmann (1990) are exceptions. But as we have indicated in the introduction, this finding is perfectly consistent with either natural monopoly tendencies in telecommunications production or an absence of such tendencies.

To illustrate this point, in Table 1 we present estimates of product-specific economies of scale and economies of scope for two studies of the Bell Canada technology that exhibit overall increasing returns to scale. These results are for illustrative purposes only, since product-specific economies of scale and economies of scope have been computed on the rather dubious assumption that the estimated cost functions could be projected into the regions of zero toll and zero local outputs. The high value of overall returns to scale (1.62) in the Kiss et al. (1981; 1983) studies is primarily due to large local-specific returns to scale, but this estimate is based on data that are now over twenty years old. Therefore, prognostication based on this evidence is not feasible. Toll-specific returns to scale are at best modest, being statistically insignificantly different from unity. Economies of scope are also estimated to be modest and statistically insignificant[24].

We present an overview of additional results in Appendix 1 before proceeding with a more detailed discussion of the recent studies that follow Evans and

[24] This cost increase of 9 percent from two-firm production is an overestimate, since it assumes that the toll-only firm must provide its own terminal equipment and local loop facilities, rather than be interconnected with the other firm's local distribution system.

Heckman. The evidence with respect to toll-specific returns to scale is very limited. The one study in Appendix 1 with an estimate is Kiss et al. (1981, 1983). The reason estimates are so scarce is that most authors recognized the difficulty of using their estimated cost functions to predict stand-alone local service costs, as is required for estimation of toll-specific returns to scale. Analogous problems exist in the estimation of economies of scope, which require the determination of stand-alone costs for both toll and local technologies. Appendix 1 contains two papers that attempt to estimate economies of scope between toll and local services (Kiss et al. [1981, 1983] and Röller [1990a]).

The subadditivity tests that were published pre-1989 all used data from AT&T. Tests using Canadian data first appeared in 1989 and will be discussed below. To our knowledge, no other country's data have been used to test for subadditivity of the cost function, with the exception of the Japanese data used by Diewert and Wales (1991b). The first U.S. tests, performed by Evans and Heckman (1983, 1984, 1986, 1988), have become very controversial. Using the local test discussed in the previous section of this paper, they rejected the hypothesis that AT&T was a natural monopoly. This finding has been disputed by Charnes, Cooper and Sueyoshi (1988), Röller (1990a,b), Pulley and Braunstein (1990), and Diewert and Wales (1991a,b), who claim to find evidence that pre-divestiture AT&T was a natural monopoly[25].

Three comments are in order. First, in our view, the major finding of these subsequent studies is not that the AT&T production technology was subadditive, but that subadditivity tests are very sensitive to imposition of regularity conditions and to changes in the functional form of the cost function. Second, as noted above, the local nature of the EH test means that this test could never confirm the natural monopoly hypothesis, only reject it. Third, the local nature of the test renders it irrelevant for the situation of competitive entry into toll production alone, since the test requires any entrant to be a 'little telephone company,' producing local as well as toll output.

Röller (1990a) is the only study of which we are aware that performs a global test of subadditivity. It is worth considering this paper in some detail, since it can be used to illustrate some of the pitfalls discussed above. The data set used is the same pre-divestiture AT&T data utilized by Evans and Heckman, extended to 1979. Röller uses two disaggregations of output: (1) local and toll, and (2) local, intra-state toll, and inter-state toll. We consider only the first disaggregation. The intra and inter distinction is not relevant outside the U.S. context, since divestiture has not been taken place in other countries. Röller finds that his estimated CES-Quadratic cost function is globally subadditive. At first glance this would appear

[25] See Banker (1990) and Davis and Rhodes (1990) for further analysis of the estimation and data problems associated with this issue. A recent paper in the genre is Wilson and Zhou (2001) which suggest that local exchange carriers costs are subadditive after introducing heterogeneity controls.

to be evidence against competitive provision of toll, since global subadditivity implies that monopoly provision of local and toll services is cost-minimizing. However, a closer look at Röller's model and results will demonstrate that this conclusion would be premature.

As observed earlier, application of the natural monopoly test to the case at issue requires projection of the estimated model into the data region of stand-alone toll production. This is a region far removed from the actual experience of pre-divestiture AT&T. Very crucially, Röller's cost function model assumes that a stand-alone producer of toll services has the *identical* dollar amount of fixed costs as a firm that produces *both* toll and local services. This assumption will bias the test towards acceptance of natural monopoly. Also, Röller's model assumes that any cost-complementarity savings associated with the interconnected nature of toll and local service facilities is lost with respect to the competitor's provision of toll. This second assumption also biases the test towards acceptance of natural monopoly.

Table 2 contains some calculations for the year in which the data are normalized (1961) and the last five years of publicly available data (1973–1977). Columns 1 and 2 contain Röller's estimates of overall economies of scale and scope. While overall returns to scale are increasing and economies of scope are positive, these results are not determinate for the question of competitive entry into toll production. They are displayed here as indicators of the general characteristics of Röller's estimated cost function.

The calculations in column 3 contain more relevant information. *SUB* (see Equation (26)) measures the degree of subadditivity in the Röller model associated with an industry structure in which pre-divestiture AT&T supplies all of the local service and 80 percent of the toll service demanded in the market, while a

Table 2
Subadditivity calculations for the Röller model

Year	Overall returns to scale[1]	SCOPE[1] (%)	SUB[2] (%)	Competitor's proportion of AT&T's total fixed costs such that SUB = 0[2] (%)
1961	1.31	17	−16	17
1973	1.30	10	−9	22
1974	1.31	9	−9	19
1975	1.32	9	−9	18
1976	1.33	9	−9	16
1977	1.34	9	−9	15

Notes: 1. Taken from Röller (1990a). 2. Calculated by the present authors.

competitor supplies 20 percent of the toll service[26]. The fact that $SUB < 0$ indicates that, according to the Röller model, our chosen industry structure would have had higher costs than a monopoly structure. During the mid-seventies competition in the form specified here would have raised industry costs by approximately 9 percent.

The Röller model assumes that the fixed costs of all firms in the industry are identical. In the present context, this means that the competitor that supplies 20 percent of the toll market and is not involved in the supply of local services has the same dollar amount of fixed costs as AT&T, which supplies 80 percent of the toll market and 100 percent of local markets. Under such conditions it is not surprising that competition is estimated to be inefficient. In column 4 we calculate the effect of moderating Röller's assumption. The numbers in this column indicate the maximum fixed cost for the competitor, as a percentage of AT&T's total fixed cost, such that the competitive structure would be no more costly than the monopoly structure (i.e. $SUB = 0$). The numbers in column 4 indicate that $SUB = 0$ when the competitors fixed costs are 15 to 20 percent of AT&T's fixed costs. Any lower percentage results in the competitive structure being cost-minimizing.

How do we evaluate this result? The Röller model implies that as long as a competitor who supplied 20 percent of the U.S. toll market (and no local service) had fixed costs no higher than 15–20 percent of AT&T's *total* fixed costs, including fixed costs associated with the provision of local service, the presence of the competitor would not have increased industry costs. This fixed cost requirement does not seem to be very stringent, and hence we would conclude that the Röller model does not provide evidence against competitive entry.

In the 1990s, economists have turned their attention to the Regional Bell Operating Companies (RBOC's). The 1984 Divestiture Decree created seven RBOC's, the holding companies that contained the 22 local operating companies of the former Bell System. Other non-Bell operating companies also existed and pre-dated 1984. These non-Bell companies ranged from larger groups such as GTE to city-specific carriers such as Cincinnati Bell[27] and Rochester Telephone. The data for these companies can prove fruitful for estimating local service returns to scale. However, many data problems exist. The source of data is the FCC's Statistics of Communications Common Carriers. Moreover, these carriers provide local services, intra-LATA toll services, as well as access for long distance (inter-LATA) calling.

Shin and Ying (1992, 1993) were the first to publish studies exploiting this data base. They used the pre-divestiture data for those RBOC's, as well as data for

[26] This division of the toll market was chosen to approximate the reality in the United States. Waverman (1989, Table 3) estimates that, as of 1986, AT&T supplied 82.3 percent of the U.S. inter-LATA (local access and transport area) interexchange market.

[27] Cincinnati Bell was no longer a Bell company at the time of the divestiture having repurchased its stock from AT&T prior to the divestiture.

non-Bell operating companies. They find modest scale economies at the sample mean[28], larger scale economies for the Bell companies[29], but no subadditivity of the cost function. They thus conclude that the break-up of the Bell System was justified on economic grounds and that competition at the local level can also be justified. The work of Majumdar and Chang (1998) suggests this conclusion indirectly. Using DEA, they calculate the returns to scale characteristics of 38 operating companies belonging to RBOC's as well as for non-Bell operating companies. They find that the majority of LEC's do not enjoy scale efficiencies, and conclude that efficiencies can be gained by an industry restructuring whereby the RBOCs are downsized while the independent operators might be allowed to expand[30].

Gabel and Kennet (1994) dispute the findings of Shin and Ying. They find positive economies of scope between switched and non-switched services, which decrease with customer density, and strong economies of scope, regardless of customer density, within the switched telephone market. Gabel and Kennet utilize an engineering optimisation model, rather than an econometric cost function. While the engineering approach permits much more attention to the details of network configuration than does the econometric approach, it does have some drawbacks. First, non-capital costs are either excluded or treated in an ad hoc manner. Typically, these costs constitute 40 to 50 percent of annual operating costs. Second, the optimisation requirement, which is central to the model, may not be reached in actual practice and there is no way to take this discrepancy into account. Third, the model assumes that the network can be configured as if there were no fixed assets currently in place. Adding assets to an existing network may have very different cost implications compared to starting from scratch[31].

Krouse et al. (1999) is a recent analysis of telecommunications technology at the local level. Their data set includes all 22 local Bell Companies that were organized into the 7 RBOCs. The time period covered is 1978–1993 period. Four outputs are considered: residential access lines, business access lines, local calls and toll calls. However, in their estimation they find that the two measures of calling have no discernible impact on costs. Thus, the only outputs considered empirically are the numbers of residential and business lines. Krouse et al. report

[28] Their estimated overall cost elasticity at the sample mean is 0.958, which implies a scale elasticity of $1/0.958 = 1.04$.

[29] Described in Shin and Ying (1992).

[30] Krouse et al. (1999) criticise Ying and Shin for not incorporating state specific regulation in their analysis. Shin and Ying, however, do include state specific dummy variables in their analysis that mitigates this issue to some extent. The period of the data set (1976–83) used by Shin and Ying predates the significant changes in regulatory oversight associated with incentive regulation.

[31] The scale economies inherent in the LECOM engineering cost model used by Gabel and Kennet can be seen by results presented in Gasmi, Laffont and Sharkey (1999) who utilize the same model. For the output range considered by Gasmi et al., the scale elasticity ranges from 2.78 to 6.25, which is considerably greater than that found in econometric studies that use historical data.

an overall cost elasticity of 1.127 at the sample mean, indicative of diseconomies of scale. They do not report estimates of economies of scope or tests of sub-additivity, since the paper is concerned primarily with the impact of divestiture and deregulation on cost efficiency.

5. Conclusions

Applied econometrics is just economic history seen through a model-oriented lens. The technical complexity of many of the econometric models of the telecommunications production technology should not obscure the fact that the empirical results can be no better than the relevance of the models' assumptions and the data used for estimation. On both counts, the econometric evidence that is reviewed in this chapter does not seem to pass the relevancy test and is not of use to economists in proposing structural changes to the telecommunications markets.

To be of relevance in these situations, the econometric models must be capable of estimating the production technology of a stand-alone toll producer. This is not the case. Nearly all of the potentially relevant models that have been estimated to date assume that the fixed costs of production are identical for a stand-alone toll producer and a producer of both toll and local services. This assumption biases the results towards a finding of natural monopoly service provision. The data sets used to estimate the models are drawn from an historical range of data far removed from the experience of a stand-alone toll producer. Extension of a cost function estimated from the available historical data to the toll-only producer's data range requires a leap of faith. That has no place in public policy proceedings.

The emerging competitive market structures will eventually provide the data necessary for the estimation of the cost functions required to test the natural monopoly hypothesis in telecommunications. The entrepreneurial experiments that competitors undertake will have an impact on the consequent evolution of industry structure in the various segments. Such experiments will generate the information that can lead to evaluations of market structure, if such evaluations are still warranted[32]. There are also new developments in estimation methods, particularly of the non-parametric variety, that are potentially useful for the evaluation of production structures of firms. Applications of these techniques to new data may shed light on the issues under contention. Until that time, econometric evidence of the type surveyed in this chapter should not, and has appeared not to, influence regulators' decisions regarding permissible market structures.

[32] By the time the required evidence becomes available, it is likely that competition will be so firmly established (at least in the U.S.) so as to be irreversible.

Appendix 1
Econometric studies of telecommunications production technology

Study	Firm	Years used in estimation	Cost model used	Outputs	Overall returns to scale (S)	MTS-specific returns to scale (S_T)	Economies of scope (SCOPE)	Sub-additivity (SUB)
Bernstein (1988)	Bell Canada	1954–78	Dynamic generalized translog	Local, Toll	1.56	Unreported	Unreported	Unreported
Bernstein (1989)	Bell Canada	1954–78	Dynamic translog	Local, Toll	0.91–2.90	n.a.	n.a.	Unreported
Breslaw-Smith (1980)	Bell Canada	1952–78	translog	Local, MTS; Other Toll	1.29	n.a.	n.a.	Unreported
Charnes-Cooper-Sueyoshi (1986)	AT&T	1947–77	Translog	Local, Toll	Unreported	n.a.	n.a.	Local: yes
Denny & others (1981a, b) Diewert & Wales (1991b)	Bell Canada	1952–76	Translog	Local, MTS; Competitive Toll	1.37–1.59	n.a.	n.a.	n.a.
1.	AT&T	1947–77	Normalized quadratic	Local, Toll	1.12–1.38	Increasing	yes	Local: yes
2.	U.S. Tel Industry	1951–87	Normalized quadratic	Local, Toll	1.26–1.68	Increasing	yes	Local: yes

continued

Appendix 1 (*continued*)

Study	Firm	Years used in estimation	Cost model used	Outputs	Overall returns to scale (S)	MTS-specific returns to scale (S_T)	Economies of scope (SCOPE)	Sub-additivity (SUB)
3.	NTT	1958–87	Normalized quadratic	Local, Toll	0.90–1.31	Constant	yes	Local: yes
Elixmann (1990)	Deutsche Bundespost	1962–86	Generalized translog	Local, Toll	0.77–1.30	Unreported	Unreported	Unreported
Evans-Heckman (1983–84; 1986 revision, 1988)	AT&T	1947–77	Translog	Local, Toll	Unreported	n.a.	n.a.	Local: no
Fuss-Waverman (1977, 1981a) Fuss-Waverman (1981b)	Bell Canada	1952–75	Translog	Local, MTS; Competitive Toll	0.89–1.15	n.a.	n.a.	Unreported
1.	Bell Canada	1957–77	Translog	Local, MTS; Competitive	1.26–1.63	n.a.	n.a.	Unreported
2.	Bell Canada	1957–77	Hybrid Translog	Toll Local, MTS; Competitive Toll	0.83–1.09	Unreported	Unreported	Unreported
Gentzoglanis & Cairns (1989)	Bell Canada	1952–86	Translog	Local, Toll	1.54	Unreported	Unreported	Local: yes
	AGT	1974–85	Translog	Local, Toll	1.06	Unreported	Unreported	Local: yes

Appendix 1 (*continued*)

Study	Firm	Years used in estimation	Cost model used	Outputs	Overall returns to scale (S)	MTS-specific returns to scale (S_T)	Economies of scope (SCOPE)	Sub-additivity (SUB)
Shin and Ying (1992, 1993)	58 LEC's and RBOC's	1976–1983	Translog	Customer access, exchange, toll merge		0.958 (local)		Not subadditive
Gabel & Kennet (1993)	n.a.	1990	Engineering cost model	Switched toll, exchange services, local private line services, toll private line service	n.a.	n.a.	Scope economies other than private line services. Diseconomies-private line and exchange services. Economies disappear in dense exchanges.	n.a.
Gasmi, Laffont & Sharkey (1997)	n.a.	n.a.	Translog on output of engineering model	CCS per customer; number of customers held constant		0.3	Unreported	Expanded Local: yes but not significant statistically

continued

Appendix 1 (*continued*)

Study	Firm	Years used in estimation	Cost model used	Outputs	Overall returns to scale (S)	Local specific returns to scale (ST)	Economies of scope (SCOPE)	Sub-additivity (SUB)
Krouse et al. (1999)	22 RBOC's	1978–1993	Translog	Intra LATA toll		Unreported		
Kiss & Others (1981, 1983) Preferred 2 output								
	Bell Canada	1952–78	Generalized Translog	Local, Toll	1.62	1.09	0.09	Unreported
Ngo (1990)								
1.	Bell Canada	1953–87	Generalized Translog	Local, Toll	1.01–2.42	Unreported	Unreported	Local: yes
2.	Bell Canada	1953–87	Generalized Translog	Local, Toll	0.75–2.05	Unreported	Unreported	Local: yes
3.	Bell Canada	1953–87	Generalized Translog	Local, Toll	0.83–1.72	Unreported	Unreported	Local: yes
Pulley & Braunstein (1990)	AT&T	1947–77	Composite	Local, Toll	Unreported but S > 1	Unreported	Unreported	Expanded local: yes; not statistically significant
Röller (1990a)								
1.	AT&T	1947–79	CES-Quadratic	Local, Toll	1.28–1.56	Unreported	0.07–0.34	Global: yes
2.	AT&T	1947–79	CES-Quadratic	Local, Intra, Inter	1.68–2.90	Unreported	0.61–0.95	Global: yes

References

Banker, R. D. 1990, Discussion: Comments on evaluating performance in regulated and unregulated markets: A collection of fables, in J. R. Allison and D. L. Thomas, eds., Telecommunications Deregulation: Market Power and Cost Allocation Issues, Westport, CT: Quorum Books.

Banker, R. D., H. Chang, and S. K. Majumdar, 1996, Profitability, productivity and price recovery patterns in the U.S. telecommunications industry, Review of Industrial Organization, 11, 1–17.

Banker, R. D., A. Charnes and W. W. Cooper, 1984, Some models for estimating technical and scale efficiencies in data envelopment analysis, Management Science, 30, 9, 1078–1092.

Berndt, E. R. and M. A. Fuss, 1986, Productivity measurement with adjustments for variations in capacity utilization and other forms of temporary equilibrium, Journal of Econometrics, October/November, 7–29.

Berndt, E. R. and M. A. Fuss, 1989, Economic capacity utilization with multiple outputs and multiple inputs, NBER Working Paper No. 2932, Cambridge, MA, April.

Bernstein, J. I., 1988, Multiple outputs, adjustment costs and the structure of production, International Journal of Forecasting, 4, 207–219.

Bernstein, J. I., 1989, An examination of the equilibrium specification and structure of production for Canadian telecommunications, Journal of Applied Econometrics, 4, 265–282.

Breslaw, J. and J. B. Smith, 1980, Efficiency, equity and regulation: An econometric model of Bell Canada, Final Report to the Department of Communications, March.

Caves, D. W. and L. R. Christensen, 1980, The relative efficiency of public and private firms in a competitive environment: The case of Canadian railroads, Journal of Political Economy, 88, 958–976.

Charnes, A., W. W. Cooper and E. L. Rhodes, 1978, Measuring the efficiency of decision making units, European Journal of Operations Research, 2, 429–444.

Charnes, A., W. W. Cooper, and T. Sueyoshi, 1988, A goal programming/constrained regression review of the Bell System breakup, Management Science, 34, 1–26.

Christensen Associates, 1985, United States postal service real output, input and total factor productivity, 1963–1984, Report to Charles Guy, Director, Office of Economics, United States Postal Service, Washington, D.C.; October.

Christensen Associates, 1990, United States postal service annual total factor productivity, report to Charles Guy, Director, Office of Economics, United States Postal Service, Washington, D.C., March.

Crandall, R. W. and L. Waverman, 1995, Talk is cheap: The promise of regulatory reform in North American telecommunications, Washington, D. C.: The Brookings Institution.

Curien, N., 1991, The theory and measurement of cross-subsidies: An application to the telecommunications industry, International Journal of Industrial Organization, 9, 73–108.

Davis, O. A. and E. L. Rhodes, 1990, Evaluating performance in regulated and unregulated markets: A collection of fables, in J. R. Allison and D. L. Thomas, eds., Telecommunications Deregulation: Market Power and Cost Allocation Issues, Westport, CT: Quorum Books.

Denison, E. N., 1985, Trends in American economic growth, 1929–1982, Washington, D.C.: The Brookings Institution.

Denny, M., M. Fuss and L. Waverman, 1981, The measurement and interpretation of total factor productivity in regulated industries, with an application to Canadian telecommunications in T. Cowing and R. Stevenson, eds., Productivity Measurement in Regulated Industries, New York: Academic Press.

Diewert, W. E. 1976, Exact and superlative index numbers, Journal of Econometrics, 4, 115–145.

Diewert, W. E. and T. J. Wales, 1991a, Multiproduct cost functions and subadditivity tests: A critique of the Evans and Heckman research on the U.S. Bell System, Department of Economics, University of British Columbia Discussion Paper No. 91–21, June.

Diewert, W. E. and T. J. Wales, 1991b, On the subadditivity of telecommunications cost functions: Some empirical results for the U.S. and Japan, Department of Economics, University of British Columbia, Vancouver, Canada.
Elixmann, D., 1990, Econometric estimation of production structures: The case of the German telecommunications carrier, Diskussionsbeiträge Nr. 57, Wissenschaftliches Institut fur Kommunikationsdienste, Bad Honnef, Germany, April.
Evans, D. S. and J. J. Heckman, 1983, Multiproduct cost function estimates and natural monopoly tests for the Bell system, in D. S. Evans, ed., Breaking Up Bell: Essays on Industry Organization and Regulation, Amsterdam: North-Holland.
Evans, D. S. and J. J. Heckman, 1984, A test for subadditivity of the cost function with an application to the Bell systems, American Economic Review, 74, 615–623.
Evans, D. S. and J. J. Heckman, 1986, Erratum: A test for subadditivity of the cost function with an application to the Bell system, American Economic Review, 76, 856–858.
Evans, D. S. and J. J. Heckman, 1988, Natural monopoly and the Bell System: Response to Charnes, Cooper and Sueyoshi, Management Science, 34, 27–38.
Fraquelli, G. and D. Vannoni, 2000, Multidimensional performance in telecommunications regulation and competition: Analysing the European major players, Information Economics and Policy, 12, 27–46.
Fuss, M. A., 1983, A survey of recent results in the analysis of production conditions in telecommunications, in L. Courville, A. de Fontenay and R. Dobell, eds., Economic Analysis of Telecommunications: Theory and Applications, Amsterdam: North- Holland.
Fuss, M. A., 1992, Econometric evidence on scale, scope and the presence of natural monopoly: What can it tell us about the desirability of competition in public long-distance telephone service? University of Toronto Working Paper No. 9210.
Fuss, M. A., 1994, Productivity growth in Canadian telecommunications, Canadian Journal of Economics, 27, 371–392.
Fuss, M. and L. Waverman, 1977, Multi-product, multi-input cost functions for a regulated utility: The case of telecommunications in Canada, paper presented at the NBER Conference on Public Regulation, Washington, D.C., December.
Fuss, M. and L. Waverman, 1981a, Regulation and the multi-product firm: The case of telecommunications in Canada, in Gary Fromm, ed., Studies in Public Regulation, Cambridge, MA: MIT Press.
Fuss, M. and L. Waverman, 1981b, The regulation of telecommunications in Canada, Technical Report No. 7, Ottawa: Economic Council of Canada.
Gable, D. and D. M. Kennett, 1994, Economies of scope in the local telephone exchange market, Journal of Regulatory Economics, 6, 381–398.
Gasmi, F., J.-J. Laffont and W. W. Sharkey, 1997, Incentive regulation and the cost structure of the local telephone exchange network. Journal of Regulatory Economics, 1997, 12, 5–25.
Gentzoglanis, A. and R. D. Cairns, 1989, Cost comparisons and natural monopoly issues for two Canadian telecommunications carriers: Bell Canada and Alberta Government Telephones, paper prepared for presentation to the Bellcore-Bell Canada Conference on Telecommunications Costing in a Dynamic Environment, San Diego.
Guldmann, J-M., 1990, Economies of scale and density in local telephone networks, Regional Science and Urban Economics, 20, 521–535.
Jha, R. and S. K. Majumdar, 1999, A matter of connections: OECD telecommunications sector productivity and the role of cellular technology diffusion, Information Economics and Policy, 11, 243–269.
Jorgensen, D. W. and Z. Griliches, 1967, The explanation of productivity change, Review of Economic Studies, 34, 249–283.
Kim, M. and J. Weiss, 1989, Total factor productivity growth in banking, Journal of Productivity Analysis, 139–153.

Kiss, F., 1983, Productivity gains in Bell Canada, in L. Courville, A. de Fontenay, and R. Dobell, eds., Economic Analysis of Telecommunications: Theory and Applications, Amsterdam: North-Holland.

Lippman, S. A. and R. P. Rumelt, 1982, Uncertain imitability: An analysis of interfirm differences in efficiency under competition, Bell Journal of Economics, 13, 413–438.

Maddala, G. S., 1977, Econometrics, New York: McGraw Hill.

Majumdar, S. K., 1995, X-efficiency in emerging competitive markets: The case of U.S. telecommunications, Journal of Economic Behavior and Organization, 26, 129–144.

Majumdar, S. K. and H-H. Chang, 1998, Optimal local exchange carrier size, Review of Industrial Organization 13, 637–649.

Ngo, H., 1990, Testing for natural monopoly: The case of Bell Canada, paper presented at the 1990 Meetings of the Canadian Economics Association, at Victoria, B.C.

Olley, R. E. and C. D. Le, 1984, Total factor productivity of Canadian telecommunications carriers, Project Report to The Department of Communications and to Members of the Canadian Telecommunications Carriers Association, Ottawa, January.

Panzar, J., 1989, Technological determinants of firm and industry structure, in R. Schmalensee and R. D. Willig, eds., Handbook of Industrial Organization, Amsterdam: North-Holland.

Pulley, L. B. and Y. M. Braunstein, 1990, A new cost function for multiproduct technologies: Specification, estimation, and applications of the composite model, Mimeo, College of William and Mary, Williamsburg.

Röller, L-H., 1990a, Proper quadratic cost functions with an application to the Bell system, The Review of Economics and Statistics, 72, 202–210.

Röller, L-H., 1990b, Modelling cost structure: The Bell System revisited, Applied Economics, 22, 1661–1674.

Seiford, L. M., 1996, Data envelopment analysis: The evolution of the state of the art, 1978–1995, Journal of Productivity Analysis, 7, 99–137.

Seiford, L. M. and R. M. Thrall, 1996, Recent developments in DEA: The mathematical programming approach to frontier analysis, Journal of Econometrics, 46, 7–38.

Shin, R. T. and J. S. Ying, 1992, Unnatural monopolies in local telephone, RAND Journal of Economics, 23, 171–183.

Simar, L. and P. Wilson, 2000, A general methodology for bootstrapping in non-parametric frontier models, Journal of Applied Statistics, 27, 779–802.

Solow, R. M., 1957, Technical change and the aggregate production function, Review of Statistics and Economics, 39, 312–320.

Uri, N. D., 2001, Technical efficiency, allocative efficiency, and the impact of incentive regulation in telecommunications in the United States, Structural Change and Economic Dynamics, 12, 59–73.

Waverman, L., 1989, U.S. interexchange competition, in R. W. Crandall and R. Flamm, eds., Changing the rules: Technological Change, International Competition, and Regulation in Communications, Washington, D.C.: The Brookings Institution.

Wilson, P., 1993, Detecting outliers in deterministic nonparametric frontier models with multiple outputs, Journal of Business and Economic Statistics, 1993, 319–323.

Wilson, W. W. and Y. Zhou, 2001, Telecommunications deregulation and subadditive costs: Are local telephone monopolies unnatural? International Journal of Industrial Organization, 19, 909–930.

This Page Intentionally Left Blank

Chapter 6

REPRESENTATION OF TECHNOLOGY AND PRODUCTION

WILLIAM W. SHARKEY*

Federal Communications Commission

Contents

* The views expressed in this chapter are those of the author and do not necessarily represent the views of the Federal Communications Commission, any Commissioners or other staff.

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.
Published by Elsevier Science B.V.

1. Introduction

As this chapter is written, in the first year of the twenty first century, the telecommunications industry is in the midst of vast structural change. The dominant regulatory paradigm is shifting from a bifurcated regime of regulation of the local exchange and competition in inter-exchange markets to one of competition, with limited residual regulatory oversight, in all markets. Both exchange and inter-exchange carriers have been involved in a round of merger activity the likes of which have not been witnessed since the early years of the previous century. Technological changes in the telecommunications equipment, information processing and entertainment industries have also been responsible for a fundamental redrawing of market boundaries, with the end result not yet in sight.

The changes occurring in the industry are happening on a worldwide scale. While the details of the regulatory environment may differ from country to country, the forces of technology and competition are the fundamental drivers of this industry transformation. The current round of structural change is only the most recent in a series of significant events in the industry throughout the previous century. In the United States, the industry has gone through a number of distinct phases. These have included: (i) a period of unregulated monopoly following the grant of the original Bell patents in the late 19th Century; (ii) the rapid entry of new entrants after the patents expired and subsequent consolidation by a dominant Bell System in the early 20th Century; (iii) a period of regulated natural monopoly under the framework of the Communications Act of 1934; (iv) the breakup of the Bell System and continued regulation of the local exchange portion of the industry in 1984 followed by the gradual deregulation of the inter-exchange portions of the industry as new entrants gained market share; and (v) a government mandate in the Telecommunications Act of 1996 to promote competition in and eventually deregulate the local exchange portion of the industry.

While the final chapter in the saga of governmental regulation of the industry has not yet concluded, it is clear that each of the previous regulatory regimes, while perhaps appropriate for its time, was ultimately undermined by the forces of changing technology and competition. These events have been surveyed elsewhere, however, and are not the subject of this chapter[1]. Instead, this chapter will focus on the technological characteristics of the industry, since a good understanding of this technology is an essential prerequisite for a deeper analysis of competition and market structure in the industry. As such, it may prove useful to both economists, who seek to explain or predict trends in the industry, and to regulators or public policymakers, who seek to influence those trends.

[1] See, for example, the chapters by Gerald Brock, David Kaserman and John W. Mayo, and Glenn Woroch in this *Handbook*.

Section 2 of this chapter provides an overview of the fundamental design principles of a telecommunications network. Section 3 describes in more detail the technologies and facilities used to provide both local and inter-city telecommunications services. Section 4 describes some recent cost models that have been used to model the technology of the local and inter-exchange networks, as well as some applications of these models in both a regulatory context and as a tool for empirical economic analysis. For the most part the technologies described in this chapter are those used to provide service on the voice grade telephone network. While data networks utilizing packet technologies are currently being rapidly deployed, and are expected to eventually become fully integrated with the public telephone network, a full treatment of these technologies is beyond the scope of this chapter[2].

2. Network design principles

2.1. Telecommunications network elements

Telecommunications can be defined as the service of enabling electronic transfers of information from one point to another or from one point to multiple points[3]. As will be explained in more detail in this and the following sections, communication involves two fundamental components: the transmission of a signal between two distinct points, and a switching function which selects a specific transmission path for the desired communication. Transmission and switching are substitute inputs in the telecommunications production function, and a well designed telecommunications network seeks to minimize cost by achieving an appropriate balance between these two functions. Point to point transmission can be accomplished along direct electrical and optical wired connections, or via wireless connections using the electromagnetic spectrum. Switching requires centralized control functions that were at one time provided by human and then electromechanical devices, but are now provided by special purpose computing machines[4].

There are important economic consequences which follow from the design of a minimum cost telecommunications network. Some portion of the cost of the network can be attributed to connecting the individual subscriber, while other costs are shared among subscribers. While a description of optimal pricing rules for telecommunications services is largely beyond the scope of this

[2] Some of the characteristics of data networks are considered in Section 3.6.

[3] Freeman (1999), pg. 1.

[4] Electronic switches have evolved over the years from stored program control analog devices, which established the connections for analog voice grade circuits, to the present day standard technology of digital switches, which allow for the switching of individual packets of information in a circuitless environment.

chapter, the cost structure of the network clearly has important implications for any pricing regime. In particular, an optimal allocation of resources will generally require that each user of the network should pay at least the incremental cost of serving its needs. Both the measurement of incremental cost and the definition of a user, whether it is a subscriber seeking access to the network, a message seeking to transit the network, or a rival network seeking access to another network, have been difficult and contentious issues facing regulators of the industry.

Economic approaches to the pricing of multiproduct technologies are now well understood[5], but the complexity of the underlying telecommunications cost function, its rapid rate of change due to technological innovation, and the distortions of pricing decisions due to the political aspects of the regulatory process, have led to a confusing and inefficient set of pricing outcomes in the industry. To the extent that prices in the industry will continue to be governed by regulatory rather than market forces, it would be desirable for regulators to make these decisions after taking full account of the technology of production.

The delivery of telecommunications services requires a set of more or less discrete network elements. The remainder of this section briefly describes their functions.

Customer Premises Equipment: Customer premises equipment (CPE) consists of the facilities on a customer's premises that allows a user to originate and terminate messages that are transmitted over the network. CPE consists of ordinary voice telephone sets, data modems, facsimile machines, computer hardware with appropriate software, PBXs and other ancillary devices such as answering machines. CPE is usually owned by the customers.

CPE was progressively deregulated in the United States during the 1970s and 1980s. In some cases, however, CPE is capable of performing functions similar to elements of the regulated public network. In particular, a PBX is nothing more than a private switching facility that allows individuals within a company, university or other organization to communicate with each other without accessing the public network. In combination with local access lines connecting the PBX to the public network, private lines connecting distant locations and, possibly, foreign exchange trunks connecting to remote switching offices, the PBX owner can create an alternative private network to the public switched network[6].

The Local Loop: Transmission facilities are the links that interconnect customer locations and switching centers throughout the network. A transmission facility connecting a customer's telephone equipment to the nearest local switching center is called a 'local loop.' The term loop is derived from the fact that these facilities

[5] See, for example, Sharkey (1982) for a discussion of pricing under complete information and Laffont and Tirole (1993) for a discussion of pricing and regulation under incomplete information.

[6] The existence of private networks creates numerous arbitrage opportunities in situations where regulated prices for interconnection or long distance usage are greater than incremental costs of these services.

have been traditionally provided by means of a pair of copper wires that form an electrical circuit from the customer's telephone to the local switch and back again[7]. Each loop is generally dedicated to the use of a particular subscriber, although loop concentration is possible when multiplexed carrier systems are used. The choice of loop technology depends on both distance and capacity considerations. For voice grade communications, a satisfactory grade of service cannot be maintained on copper wires beyond a distance of approximately 15,000 feet to 18,000 feet unless such elements as 'loading coils' are used to maintain signal quality in the mid-band voice frequency range. Other services, including both residential and business data transmissions, generally require a shorter copper distance in order to maintain satisfactory service quality. Loop carrier systems are used both to improve transmission (shorten the length of copper facilities) and to concentrate (reduce the amount of copper) loop facilities. Loop carrier may utilize either copper or fiber media. As the price of electronics continues to decline, local carriers are pushing fiber closer to the subscriber.

Interoffice Trunks: Transmission facilities that connect switching centers are referred to as 'trunks.' For very short interoffice links, bundles of wire pairs can be used. For the greatest part of the interoffice and long distance network, however, high capacity optical fiber or copper carrier systems are used. The original carrier systems were analog systems based on frequency division multiplexing[8]. Beginning in the 1960s digital carrier systems were introduced based on time-division multiplexing[9]. Carrier systems were originally provided on copper pairs. Since their introduction, advances in technology have introduced co-axial cable, terrestrial microwave, satellite, and optical fiber technologies for intermediate and long haul transmissions. Optical fiber systems, and for some transmission links satellite transmission systems, are now the predominant media used in high capacity trunking systems.

Tables 1 through 3 illustrate the pace of technological advances in transmission technologies during the second half of the 20th century.

During the 1980s the fastest coaxial carrier systems were able to transmit approximately 8×10^8 bits per second. Current transmission technologies rely primarily on fiber optic media based on the Synchronous Optical Network (SONET) and other standards. SONET bandwidths are expressed as optical

[7] Cable pairs typically consist of 19. 22. 24 or 26 gauge copper wires that are twisted so as to minimize coupling with other circuits. Pairs are bundled into binder and cable groups of from 6 to 4,200 or more pairs.

[8] In frequency division multiplexing, multiple voice grade channels are transmitted together by keeping them separated in frequency. Multiplexing and de-multiplexing electronic devices are used to convert the voice grade frequencies into frequencies that can be simultaneously transmitted (Bell Laboratories [1980], p. 135–40).

[9] In digital time division multiplexing, each voice frequency signal is sampled and converted into a binary code of 0s and 1s. The pulse streams from a number of signals are then transmitted on the same transmission medium by interleaving these pulse streams (Bell Laboratories [1980], p. 315–20). 24 voice grade channels can be carried on a T1 system and 672 channels can be carried on a T3 system.

Table 1
Digital transmission systems

Digital service level	Voice grade equivalents	Line rate[10]
DS-0	1	64 kbps
DS-1	24	1.544 mbps
DS-3	672	45 mbps

carrier rates which currently support transmission levels from OC-3 and higher rates[11]. Looking forward, optical fiber technologies have been deployed having a carrying capacity of more than a terrabit per second (1×10^{12} bits per second)[12]. These systems use wave division multiplexing to combine multiple colors of light with each color carrying OC-48 or higher on one fiber pair.

Switching: The purpose of switching is to interconnect transmission paths, which makes it possible and economically feasible for point to point communications to be made throughout the network without the need to establish a direct transmission link between every pair of points. The first switch consisted of a manual switchboard in which a human operator made connections based on verbal instructions from calling parties. Various automatic switching machines were introduced during the early 1900s including step-by-step, panel and crossbar designs (Bell Laboratories, 1980, pages 236 to 45). These innovations introduced common control features to switching machines in which switching functions were disaggregated and placed under a controlling device that interpreted the caller's dialing code. This design allowed for a more efficient utilization of switching capacity, since the individual functions could be directed to new tasks before a previous call setup was completed[13].

More recently, control features have been implemented in software programs, resulting in cost savings associated with adding new customers or new features to the switch. Stored program control switches were first based on analog 'space-division' methods, in which all internal connections were accomplished through direct metallic contacts or semiconductor cross-point devices. Typical telephone

[10] DS-0 at 64 kbps results a sampling rate of twice the highest frequency sampled. This results in a rate of 2×4000, and with 8 bits per sample this equals 64,000 kbps.

[11] In the SONET standard, bandwidths begin where previous digital transmission technologies reach their upper limits. An OC-1 transmission level corresponds to a DS-3 level of approximately 51 Mbps. Outside of North America, a closely related system known as the synchronous digital hierarchy (SDH) is used.

[12] Wildeman and Flood (2000, p. 11). Current SONET transmission rates extend to OC-192 and higher levels.

[13] The migration of switching technologies toward greater utilization of common control features has also shifted the cost structure of switching by increasing the initial fixed cost of the switch and reducing the usage sensitive variable cost.

Table 2
Voice channel capacity of coaxial and microwave transmission media*

Year of introduction	Voice grade equivalents	Technology
1946	1,800	L1 Coaxial
1950	2,400	TD-2 Microwave
1953	9,300	L3 Coaxial
1967	32,400	L4 Coaxial
1968	12,000	TD-3 Microwave
1973	16,500	TD-3A Microwave
1974	108,000	L5 Coaxial
1978	132,000	L5E Coaxial
1981	42,000	AR-6A Microwave

* Source: Minoli (1991, p. 3).

switching technologies now utilize digital time-division switching methods (Bell Laboratories, 1980, p. 270–3). In addition to the advantages of continuing cost reductions brought about by technical advances in computing facilities generally, digital switching technologies allow for a seamless connection with digital transmission plant without the need for digital to analog conversions. In a further evolution of switching technologies, some equipment vendors have proposed a 'softswitch' concept under which call processing and physical switching functions are completely separated, while connected by standard protocols. An advantage of this approach is the potential for cost reductions in switching as new entrants with specialized expertise in call processing features are able to enter the market. The establishment of standard protocols should also allow for a seamless migration of switching equipment from circuit based to packet technologies.

Table 3
SONET transmission standards*

Standard	Line rate
OC-1	51.84 Mbps
OC-3	155.52 Mbps
OC-12	622.08 Mbps
OC-48	2.48832 Gbps
OC-192	10 Gbps
OC-768	40 Gbps

* Source: Wright (1993, p 40) with updates.

Signaling: Signaling functions are closely related to the switching element, and are essential for call setup and other network control functions. Local signaling consists of the provision of dial-tone on demand, receipt of the dialed digits and provision of the ringing signal which is sent to the called party's receiver (Bell Laboratories, 1980, p. 174–5). Interoffice signaling consists of the machine to machine messages which allow the local switching offices of the calling and called parties, as well as any intermediate switching centers, to process the call. Current interoffice signaling technology, known as common channel signaling (CCS), uses a signaling path that is independent of the transmission path used for the call itself. CCS makes use of a separate high speed packed switched network to propagate switching commands throughout the network. CCS is an 'out-of-band' signaling system that substantially improves security and the efficient utilization of the network. This architecture also provides significant advantages as telecommunications carriers increasingly offer a mix of data and voice capabilities.

2.2. *The trade-off between switching and transmission*

One of the most basic examples of telecommunications network planning is the use of switching capacity to substitute for transmission capacity. By applying more complex routing algorithms within switching machines, and making use of alternative routing possibilities within the network, economies of scale in individual transmission links can be taken advantage of and overall network costs can be reduced. While the full-scale network optimization problem is complex for even moderately large networks, the basic principles involved are very simple. Consider, for example, a network consisting of 8 customer nodes located on a regular grid as in Figure 1. In Figure 1a, a 'mesh' network, consisting of direct transmission paths between every pair of customers is illustrated. In this example the total length of transmission links is 95.51 distance units[14]. The number of links required is $n(n-1)/2$ or in this example 28. In Figure 1b, a 'star' network is constructed by locating a single switch at a central location and connecting every customer directly to the switch. In this example, the number of links reduces to 8 (28 percent of the previous total) and the total length of transmission links decreases by more than 90 percent to 9.15 distance units. Of course, the cost of the overall network must take account of the cost of switching as well as transmission, but one should keep in mind that both the star and the mesh network require switching capacity. The difference is that the switching capacity is centralized in Figure 1b, while in Figure 1a, switching must be provided at each customer location.

In Figure 1c an additional switch is added, which allows for a further substitution of switching capacity for transmission capacity. The total transmission

[14] The distance between two adjacent customers in both a vertical and horizontal direction is assumed to be one unit.

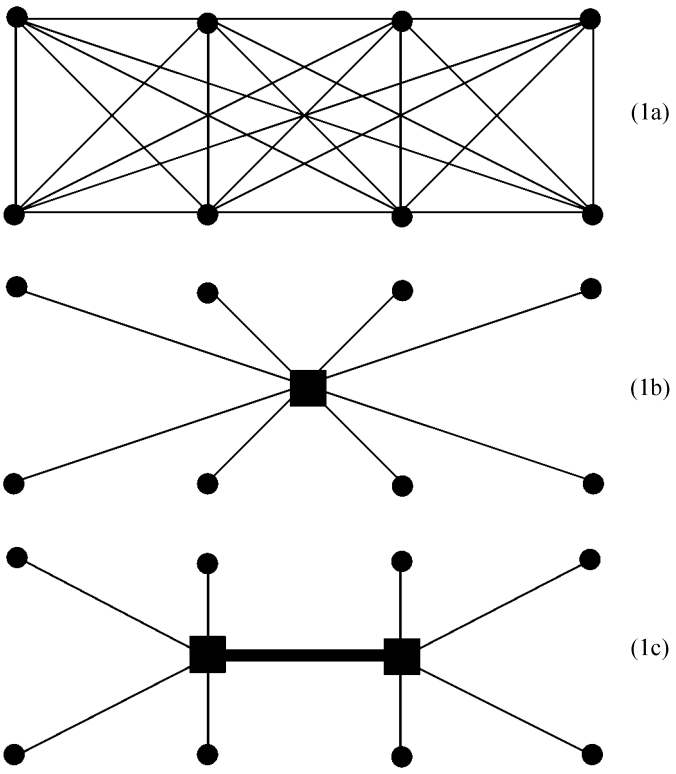


Fig. 1. The trade-off between switching and transmission.

distance in this network is now 7.47 distance units. If the original eight nodes all represent customer locations, then Figures 1b and 1c represent two different feasible local exchange networks that could be constructed to serve these customers. In the network of Figure 1b, a single local exchange serves all eight customers; while in the network of Figure 1c, there are two local exchanges, each serving four customers.

The lines connecting the switches are called interoffice trunks. In evaluating the relative efficiencies of the two possible networks, several factors must be taken into consideration in addition to the relative costs of raw switching capacity versus raw transport distance. First, there are typically economies of scale in transmission facilities; as a result, the interoffice trunking facilities will have a lower cost per circuit mile than the loop facilities. In addition, trunking facilities are a shared resource, since it is unlikely that each of the four customers served by one of the local switches would wish to place a call to a customer on the second switch at exactly the same moment in time. Because of the

sharing of the trunking capacity, far fewer transmission channels are required between the two switching centers than the total number of customer lines served by those centers.

2.3. Telecommunications traffic theory

This section describes the basic tools that are used to determine optimal capacity on traffic sensitive transmission links. In the local loop, several subscribers may share the cost of a particular link in the network, but the network optimization problem in designing local facilities is concerned only with the number and location of subscribers, and the number of communications channels that each subscriber demands. In the interoffice and intercity networks, individual communications channels are not, in general, dedicated to particular users. The task of the telecommunications engineer is to determine the required capacity of each link based on expected traffic demands that the population of users as a whole place on the network[15].

As with most network optimization problems, a full solution to the traffic-engineering problem is extremely complex, and in most cases complete analytical solutions are not possible. With some simplifying assumptions, however, relatively simple models can be constructed that capture the essential features of the problem, and in some cases these models offer good approximations to the best available solutions. The classes of models that are used by traffic engineers are models of queuing systems (also called stochastic service systems). In these models, there is a set of users who send streams of orders to a service center. Both the arrival rate and the processing rate of each type of order are treated as random variables. The service system is composed of a fixed number of servers, i.e. communications channels, and a service or queuing discipline that determines the order of service of calls and the fate of unserved calls. If all available servers are busy when a new order arrives, the service discipline may allow arriving orders to wait in a queue for the next available server. If orders arrive when there is inadequate buffer space they are blocked. Given appropriate assumptions about call arrival rates and holding times, the probabilities of blocked calls and call delay can be calculated. The objective of the traffic engineer is to determine the number of communications channels in order to maintain the probabilities of blocking or waiting within tolerable quality of service limits[16].

Formal analysis of telecommunications service systems is attributed originally to A. K. Erlang. The Erlang model assumes that calls arrive according to a

[15] In planning for interoffice capacity, some circuits are dedicated to individual users. These demands are simply aggregated with the shared capacity needs in determining total network capacity.

[16] Closed form solutions to the relevant probabilities can be easily obtained in the case of a single server and a zero or infinite buffer capacity. For multiple servers and a positive but finite buffer, traffic engineers generally rely on numerical techniques to provide approximate solutions.

stationary Poisson process at rate λ[17]. The probability density function for arrivals at instant t is given by $\lambda e^{-\lambda t}$. Holding times for calls are assumed to be independent and exponentially distributed with a mean holding time $1/\mu$, where the parameter μ represents the service rate. The probability density function for holding time is given by $\mu e^{-\mu t}$. In the Erlang blocking model, it is assumed that there is no buffer capacity so that calls which arrive when all circuits are busy are cleared from the system. The traffic load is given by $a = \lambda/\mu$, and represents the average number of servers that could be kept fully occupied by incoming calls. A value of $a = 1$ represents a dimensionless traffic intensity of one Erlang. In the United States, traffic loads are typically expressed in units of hundreds of call seconds, or CCS (for *centum* call seconds per hour). Since there are 3,600 seconds per hour, it follows that one Erlang = 36 CCS.

When there are n available servers, it can be demonstrated that the probability that a call is blocked, i.e. that it arrives when all n servers are busy, is given by[18]

$$B(n, a) = \frac{a^n/n!}{\sum_{k=0}^{n} a^k/k!}$$

In a blocking system, in which blocked calls are lost, the quality of service can be expressed in terms of the blocking probability. Generally, a blocking probability of $\beta = 0.01$ is considered satisfactory. If $G(n)$ represents the cost of providing n communications channels (servers) then the traffic engineer's problem can be represented by the minimization problem.

$$\begin{aligned} &Min \quad G(n) \\ &n = 0, 1, \ldots \\ &subject\ to\ B(n, a) \leq \beta \end{aligned}$$

It has been noted previously that the transmission cost function $G(n)$ typically exhibits significant economies of scale. When costs are computed as a function of offered loads, an additional source of scale economy is the found in the probabilistic nature of traffic. For example, the solution of the traffic engineer's problem gives rise to a trunk requirements function as illustrated in Figure 2. In this example, 2 trunks can carry a load of 0.15 Erlangs; 4 trunks can carry a load of 0.87 Erlangs; 8 trunks can carry 3.13 Erlangs; and 16 trunks can carry 8.88 Erlangs.

[17] A Poisson process is characterized by the property that the intervals between two adjacent arrivals are independent random variables with identical exponential distributions. Such a process has no memory, so that the probability of an additional arrival during any interval of time does not depend on the current state of the system. In a stationary process, the expected number of arrivals during any interval does not depend on the time at which the interval begins.

[18] For a derivation, see Feller (1957, pages 413 to 421).

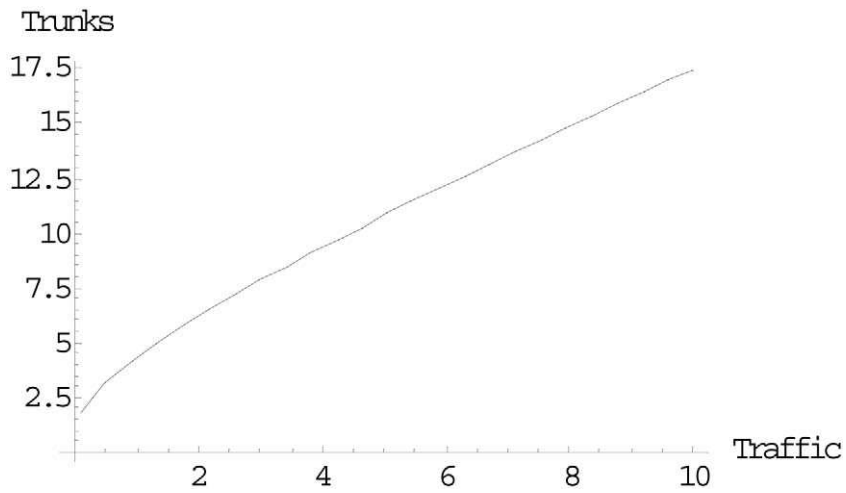


Fig. 2. Trunks required to maintain service quality $\beta = 0.01$.

In some telecommunications applications, the quality of service depends on waiting time in the service system instead of the probability of a blocked call attempt. For example, in a common control switch, the control device receives the set of calling digits from the person initiating a call and stores this information while a conversation path is established in the network. When the control device is busy serving other requests, the calling party will experience a delay in call setup. Switching equipment is therefore designed so as to minimize the probability of excessive delay[19]. Similar issues arise in the design of broadband networks, where delay rather than blocking of individual packets represents a more reasonable performance criterion.

2.4. Optimization in the local access network: Minimum distance and minimum cost networks

This section describes some of the methods and results that network planners, and modelers, use to design a telecommunications network. In the local access network, a minimum cost network can typically be represented as a 'spanning tree' network. Interoffice and intercity networks generally require a more complicated network structure, since the volume of traffic flows must be considered as well as the costs of establishing simple connectivity. For a technical

[19] For example, switches may be designed so that the probability of a delay in call setup greater than 3 seconds during the busy hour is less than one percent. This probability can be expressed in terms of the Erlang delay function which is derived in Bell Laboratories (1980), p. 492. Absent network abnormalities, the probability of delay outside the busy hour is far less than one percent.

graph theoretic presentation of some of the basic issues involved in network optimization, the reader should refer to the recent surveys contained in Ball et al. (1995a, 1995b). This section will describe some of the main results in a non-technical fashion that are most relevant to the design of telecommunications networks.

A classic network optimization problem is the minimum cost network flow problem, which consists of a set of *nodes* and a set of directed *arcs* (pairwise links connecting nodes). Each arc has an associated cost as a function of the flow demand carried on that arc. The general objective function in such a problem is to minimize the cost of transporting a particular set of flow demands through the network[20]. In the case of a local access network, which connects a set of subscribers to a switching location, the flow demands represent demand for a dedicated capacity which allows the user to access the switch. In the case of an interoffice network, the flow demands represent individual message units, measured in minutes of voice traffic or packets of data traffic, transmitted throughout the network.

In the case of local access networks, optimal solutions can be obtained relatively easily as a result of following 'spanning tree property,' defined below. A spanning tree is a minimal connected network, i.e. there is a path between any two nodes in the network, but if any link in a spanning tree network is removed, the network is broken up into two disjoint parts[21].

> *Spanning Tree Property:* A concave cost network flow minimization problem whose objective function is bounded from below over the set of feasible solutions always has an optimal spanning tree solution. This is an optimal solution in which only the arcs in a spanning tree have positive flow.

Telecommunications transmission technologies generally satisfy the concavity property. This is equivalent to the property that the average cost of a flow declines as the flow capacity increases. In the case of a local access network, flow demands represent additional access lines, and the fixed costs associated with structures such as telephone poles and underground conduits are sufficient to guarantee the concavity of the arc cost functions in this context[22].

Computation of minimum cost spanning trees turns out to be extremely simple, and computationally efficient, based on various algorithms. One explicit algorithm was discovered by Prim (1957). A minimum distance net-

[20] See Ball et al. (1995a, p. 3).

[21] Alternatively, a spanning tree network can be defined as a graph in which there are no cycles – i.e. no paths are created which start from a given node and return to that same node following links in the network.

[22] Arc cost functions in the flow problem relevant for interoffice capacity planning are also generally concave. However, the flow demands in this context are general point to point flows between offices rather than flows from subscribers to a common switch, and the spanning tree property no longer holds in this situation.

work can be constructed using the Prim algorithm in the following way. Beginning with a network consisting only of the central office, find the nearest customer node that is not yet attached to the network and attach it. The network then consists of the switch and one customer. The algorithm proceeds step by step in attaching customer nodes to the network on the basis of minimum distance. At any point in the algorithm, it chooses from the set of all nodes that have not yet been attached a node that is closest to some node in the existing network. Prim demonstrated that this simple algorithm, an example of a class of 'greedy' algorithms, necessarily leads to a minimum distance network. In other words, when the algorithm is completed, there is no possible way to reconfigure the network so as to lower the aggregate distance of all links in the network.

As long as structure costs are significantly larger than cable costs, the Prim algorithm provides a satisfactory solution to the design of minimum cost local access network. In moderate or high-density communities, however, a minimum distance spanning tree network is not generally optimal. While it minimizes the total distance of all links in the network, it does not minimize the distance between any particular node and the switch. If a particular customer (for example, a large business user) requires a large number of access lines, then an efficient network design may require that the path for that customer to the switch should be minimized even if the total distance of the network increases.

There are no simple algorithms that can be applied to give a completely optimal solution to the general problem. (A "star" network, which minimizes the cost of connecting each node to the central office, would not be optimal because it does not take advantage of potential sharing of structure costs in the network). However, it is possible to modify the Prim algorithm to take account of the effects of traffic in the network on total cost, and generally to improve the performance of the algorithm computationally.

A minimum distance spanning tree network may also fail to be a minimum cost network for another more subtle reason. The minimum distance-spanning network is constructed using only the links connecting pairs of customer nodes or pairs connecting a customer directly to the switch. However, if it is possible to create new 'junction' nodes it may be possible to reduce the cost of connecting the existing set of customer nodes to the switch[23]. Figure 3 illustrates this possibility in the case of a simple triangular network. If two customers and a central office are located at the vertices of an equilateral triangle, the minimum distance spanning tree network would use two sides of the triangle as links. Total distance could be reduced by adding a junction node at the center of the triangle. In this case the addition of junction nodes reduces total network

[23] By the same logic, it is possible that the addition of a new customer to a network can actually reduce the total cost of connecting all customers to the switch if junction nodes are not allowed.

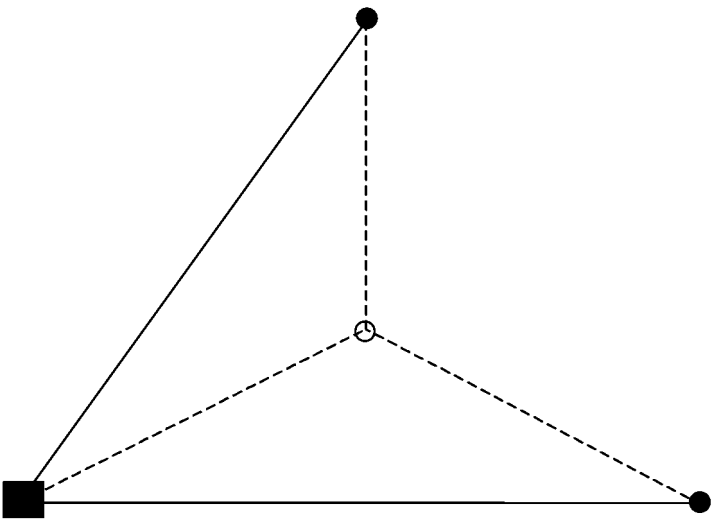


Fig. 3. A network with a junction node.

length by a factor of 13 percent compared to the length of the spanning tree network[24].

2.5. Network optimization in the interoffice and inter-city networks

The use of switching centers to concentrate traffic into high capacity trunking groups is a design principle that is applied throughout the interoffice and inter-city network. An additional design principle, that further increases the concentration factor for high usage trunk groups, is the principle of alternate routing. Alternate routing methods are both static and dynamic in nature. Under static routing, the objective of the network engineer is to achieve an optimal trade-off between the length of a route between any two points and the economies of scale that can be achieved by aggregating traffic. As noted previously, telecommunications transmission media are subject to strong economies of scale. Therefore, in many cases a longer indirect route, which allows capacity to be shared with other point to point traffic, can achieve lower costs than a direct route.

Dynamic routing methods also take account of the time profile of demand. These methods can be both time invariant and time varying. Using time invariant routing methods, it may be efficient to route some point to point traffic along a direct link during off-peak periods, while overflowing the additional peak period

[24] The network with junction nodes has a length which is $\sqrt{3/2}$ time the length of the spanning tree network. It has been demonstrated that adding junction nodes can never decrease the length of a spanning tree network by more than this amount.

traffic onto an indirect route. These approaches can take advantage of the regularity of calling patterns by time of day which is frequently observed due to the differing concentrations of business and residential customers throughout the network. Under a time varying routing method, traffic measurements would be taken continuously during the day, and routing methods selected in real time. By taking advantage of these non-coincident busy hours, facilities that otherwise would be underutilized can be used to delay or eliminate network builds.

In simple network situations, it is possible to illustrate the cost savings possible due to alternative routing methods. In designing an efficient interoffice network, engineers must first estimate the expected peak period traffic between each pair of offices. If traffic demand is sufficiently large, then a direct trunking group may be called for, as illustrated by the link between nodes A and B in Figure 4. When the volume of traffic between two switching centers is lower, then these demands can be pooled with other pair-wise interoffice demands by routing the traffic through an intermediate "tandem" switching location. For example, in Figure 4, traffic from nodes A to C is routed through a tandem switch T. The link between A and T carries both the traffic from A to C and the traffic from A to other nodes such as D. Similarly, the link between C and T might carry traffic between node C and nodes A, B and D.

As an illustration of time varying routing, it may be efficient to carry a base load demand on a direct trunk group, such as the link from A to C in Figure 4. If the capacity of this route is designed to be less than the expected peak period traffic between nodes A and C, then the overflow traffic can be routed through the tandem switch T. In the intercity toll network these principles are implemented by means of a multilevel hierarchy of switching centers. These switching centers have detailed routing tables which define the path that a particular call

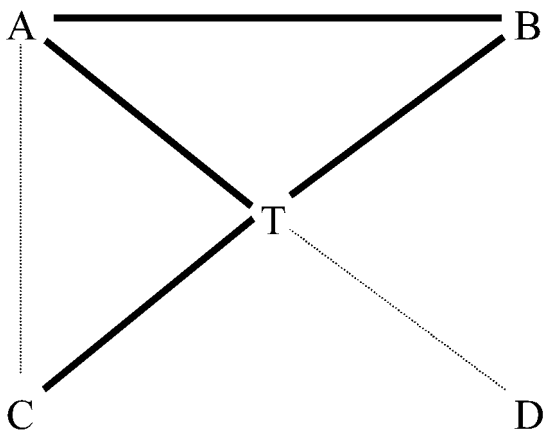


Fig. 4. Alternative routing.

should take based on the facility utilization at any point in time. In some cases, aggregate traffic in a region may exceed the traffic capacity of a single tandem switch. Hence, either additional direct paths or additional tandems must be added, and an additional hierarchical layer of routing decisions must be made[25].

3. Network facilities and technologies

A telephone network must allow any customer to connect to any other customer. In order to accomplish this task, a telephone network must perform three functions. The network must connect customer premises to a local switching facility; and it must ensure that adequate capacity exists in that switching facility to process all customers' calls that are expected to be made at peak periods. It must then interconnect that switching facility with other switching facilities to route calls to their destinations[26]. The purpose of this section is to describe the network facilities and technologies that are used to accomplish each of these tasks.

3.1. Wireline subscriber access

Figure 5 provides a visual representation of a wireline local access network. In a wireline local loop, the cables, structures and other equipment that connect the central office to the customers' premises are known as outside plant. Outside plant can consist of either copper cable or a combination of optical fiber and copper cable, as well as associated electronic equipment. Copper cable generally carries an analog signal that is compatible with most customers' telephone equipment[27]. The range of an analog signal over copper is limited, however, so that thicker and more expensive cables or loading coils must be used to carry signals over greater distances[28]. Optical fiber cable carries a digital signal that is incompatible with most customers' tele-

[25] For a more extensive discussion of alternative routing principles see Bellcore (1994, Section 4, p. 17–40).

[26] For simplification, the terms local switching facility and wire center are used interchangeably in this chapter. In a modern network, they often are not at the same physical location. A wire center is a point of concentration of local network facilities. Most wire centers now in service were established before the advent of either loop carrier systems or digital switching. Both of these technologies likely would dramatically alter the trade-off between switching and transmission technologies that existed when wire centers were first established. The likely result would be far fewer local switches than are currently deployed by domestic local exchange carriers.

[27] For all copper loops, copper feeder connects from the main distribution frame (MDF) in a wire center to a feeder distribution interface (FDI) in the local network. Copper distribution cables extend from the FDI toward customer premises.

[28] Since 1980, following the wide spread availability of digital loop carrier (DLC) systems, local loops have been designed using only finer gauge wire (24 or 26 gauge, not 19 or 22 gauge) and no load coils. Initially, DLC used T-1 carrier over copper wires to connect the remote DLC location to the wire center. Most modern DLC utilizes optical fiber based carrier.

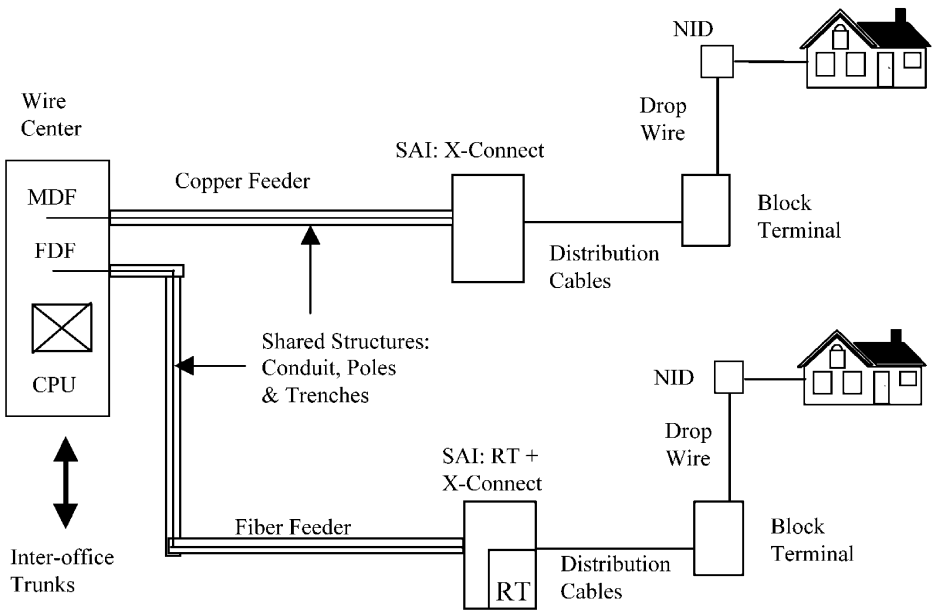


Fig. 5. Wireline local access.

phone equipment, but the quality of a signal carried on optical fiber cable is superior at greater distances when compared to a signal carried on copper wire.

Generally, when a neighborhood is located too far from the wire center to be served by copper cables alone, an optical fiber cable will be deployed to a point within the neighborhood. At that interface point, called a serving area interface (SAI), a piece of electronics equipment will be placed that converts the digital light signal carried on optical fiber cable to an analog electrical signal. This equipment is known as a digital loop carrier remote terminal, or DLC. Fiber cables connect the fiber distribution frame (FDF) at the wire center to the DLC terminal at the SAI. From the SAI, copper cables of 24 or 26 gauge extend to all of the customer premises in the neighborhood. Where the neighborhood is close enough to the wire center to be served entirely on copper cables, copper trunks connect the main distribution frame (MDF) at wire center to the SAI, and copper cables then connect the SAI to the customers in the serving area. The portion of the loop plant that connects the central office with the SAI is known as the feeder plant. Distribution plant runs from the SAI to a block terminal near each customer's location. A drop wire[29] from the block terminal connects to a network interface device (NID) at each customer's location. As the cost of fiber electronics decreases, LECs are extending fiber optics ever closer to the customer premises.

[29] In multi-tenant buildings riser cable (inter-building cable) often replaces the drop wire.

3.2. Alternative subscriber access technologies

Cellular radio systems allow for the efficient use of the radio spectrum for providing an alternative method of local access to the telecommunications network (see the chapter by Hausman in this *Handbook*). While primarily used for mobile access uses, fixed cellular technologies also have a potential use in providing low cost access in areas in which the density of subscribers is low. This section also considers cable access and all-fiber access technologies.

Cellular Mobile Access: Cellular service works by dividing a service area into a large number of individual cells, with the shape of a hexagon. Each cell has a radio tower that transmits and receives radio signals to and from portable telephone equipment using the minimum power output that is able to cover the entire cell. The radio frequency used for a given cell can, therefore, be reused except in cells that directly border the original cell. The key technology that makes cellular possible, however, is the switching technology that allows a radio transmission to be handed from one cell to another in order to follow a moving object. As switching technologies become faster and more reliable, it is possible to decrease the size of the cells, and, therefore, to increase the capacity of the service, which is limited by the available quantity of radio spectrum available. The mobile switching center performs all call setup, switching and signaling facilities that are needed to allocate spectrum and other shared links within the network to individual subscribers for the duration of a call. Generally, a mobile switching center performs all of the functions of both a local office and a tandem switch in a wire line network.

Cellular systems were originally based on analog transmission technologies, but as in wireline transmission technologies, digital technologies have increasingly been deployed in ways that offer significantly more efficient use of the spectrum. Both time division multiplexing and code division multiplexing technologies are currently being deployed. Subscribers to a wireless network share spectrum, and the common control facilities that assign spectrum to individual users for the duration of a call are an integral part of the 'wireless local loop.' Accordingly, a significant part of the costs of connecting a wireless subscriber to a point of interconnection may be considered as traffic sensitive. The fact that none of the costs of a wireless network can be identified with dedicated subscriber lines, however, does not imply that the entire cost of a wireless network should be considered traffic sensitive. Conceptually speaking, there remain potentially significant common costs that are not sensitive to additional traffic.

Fixed Wireless Access: A fixed wireless access network follows the same design principles as a mobile network, except that the cost of access can be reduced by using directional transmission technologies. In addition, smaller cell sizes can be used since the switching software is not required to service mobile users. These factors can significantly reduce the traffic sensitive costs, while possibly increasing

the fixed costs. A fixed wireless access network should also expect to plan for usage levels that would be comparable to usage in a wire line network. A 1994 study by two consulting firms assumed a busy hour utilization figure of 0.1 erlangs (3.6 CCS) for a fixed wireless system and 0.02 erlangs (0.72 CCS) for a cellular system[30].

Cable Television Access: Cable television networks have been traditionally designed for broadcast transmission of a common high bandwidth signal to all customers on the network. The architecture of a cable distribution network is superficially similar to that of a wire line telephone network, as illustrated in Figure 5, except that the distribution portion of the network generally uses coaxial cable instead of copper pair technologies. A more subtle difference between the two technologies is the need for signal amplification. Cable television signals occupy significantly more bandwidth than voice telephone signals, and travel is in one direction from the headend, comparable to a local office, to the subscriber, rather than bi-directional. Amplifiers are required to maintain signal quality for the broadband signals, but they also degrade the quality of the signal and the reliability of the service, since the failure of a single amplifier can disrupt service to all customers who are located downstream from the unit that fails. Fiber feeder cable also is increasingly being deployed in cable networks.

Until recently, the primary difference has been that the wire line network creates a dedicated, narrow-band, two-way link between the customer and the serving wire center, while the cable network creates a shared, one-way, broadband link. Wire line networks are increasingly making use of copper DSL technologies in the distribution portion of their networks to extend broadband over copper. Similarly, some cable networks are adding two-way broadband and narrow band capabilities. On a forward-looking basis there would appear to be few significant differences between cable access and wire line telephone access technologies.

Fiber Access Technologies: As described in Section 3.4 below, fiber rings are currently an efficient technology for the interconnection of switching centers. For large users, the same technologies have been used to directly link large customers to a local or tandem switch or an inter-exchange carrier's point of presence. In spite of the rapid growth in the demand for broadband services by residential customers, however, fiber technologies are not currently being deployed for residential use. 'Fiber to the home' technologies are expensive for two reasons. First, a fiber signal to the home requires a remote terminal at the customer's premise to convert the optical signal to an electronic signal which can be recognized by the customer premise equipment. Second, fiber distribution technologies require a supplementary electrical source to supply power for signaling and ancillary services.

[30] Economics and Technology, Inc./Hatfield Associates, Inc. (1994, page 78).

3.3. Switching

Telecommunications switches are input-output devices that select the transmission paths that a given call will take throughout the network. Current electronic switching machines use both space division and time division methods to achieve good performance at minimum cost[31]. All modern switches are stored program control switches. These switches contain the following elements: (1) the switching matrix representing the electromechanical or solid-state components where the input-output connections are made; (2) the call store, which consists of temporary memory storage for incoming call information and availability of outgoing circuits; (3) the program store containing permanent memory with instructions and translation information for the central controller; and (4) the central processing unit[32].

Switches may involve concentration of traffic. For example, a local end office switch requires an input connection to every subscriber in the wire center served by that switch. Generally, on the output side fewer ports would be needed. This difference between the number of line ports and trunk ports is a common measure of concentration. A tandem switch, on the other hand, would have similar numbers of trunk ports on both the input and output side of the switch, and would not necessarily be designed to concentrate traffic. Switch ports represent a fixed, non-traffic sensitive, cost of the switch. Of the remaining switch components, the capacities of the switching matrix, and to some extent the memory stores, depend on the peak period traffic for which the switch is designed. The central processor is largely a non-traffic sensitive cost component.

Many of the non-traffic sensitive costs can be eliminated in a remote switch. This is sometimes called a satellite exchange. Remote switches have true switching functionality in that calls which originate and terminate in the same satellite serving area can be processed entirely by the remote unit[33]. However, for calls which originate or terminate outside the satellite area, the remote unit must communicate with the parent (host) switch in order to function. Both remote switching units and digital loop carriers in remote terminals are used to reduce the cost of serving sparsely populated rural areas.

In the inter-exchange portion of the network, tandem switches are generally deployed in a multi-level hierarchical arrangement in circuit switched networks. These networks rely on alternative routing methods, discussed in Section 2.5, to minimize the cost of the overall network. Packet networks, including internet backbone facilities, rely on different switching technologies using TCP/IP or other protocols. Each packet contains headers which give the destination address and

[31] See Freeman (1999, pp. 133–142) for a discussion of some of the design principles for digital electronic switches.

[32] Freeman (1999, p. 78).

[33] In contrast, a digital loop carrier located at a remote terminal described in the section on wire line access technologies, is purely a concentration or analog to digital conversion device that does not have any switching capability.

information which allows all transmitted packets to be reassembled at the distant end in the correct order, even though individual packets are likely to travel on different paths to their destination. The protocols also provide error-checking routines which automatically re-transmit packets that are lost or corrupted during transit.

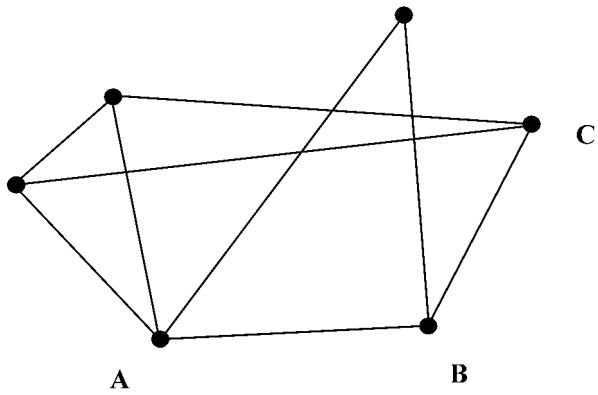
3.4. Interoffice transport facilities

Depending on traffic intensities and transmission costs, any interoffice network design, from a tree to a full mesh network, can be cost minimizing. In the network of Figure 6a, traffic between many office pairs, for example, from A to B and from B to C, is carried on direct trunk groups. Other traffic, such as traffic between A and C is carried indirectly through one or more intermediate offices. As the ratio of variable cost to fixed cost on individual transmission links decreases, the minimum cost interoffice network tends to become more sparse, having fewer direct trunks between office pairs and placing a greater reliance on indirect routing. Rapid technological advances in fiber transmission technology have created a situation in which very few office pairs can justify a direct trunk connection for short haul interoffice traffic. Many interoffice networks, including networks of host and remote switches and networks connecting end office switches to nearby tandem switches, are now designed on the basis of a ring topology as in Figure 6b[34].

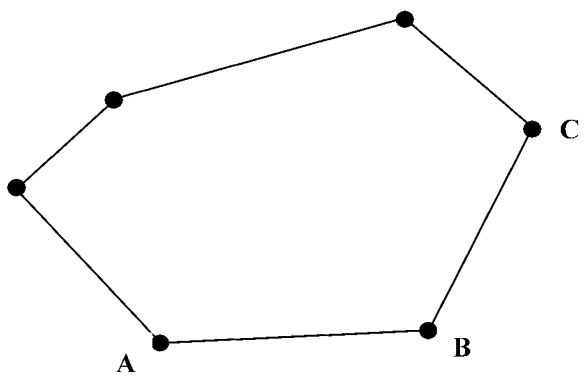
Based on a Synchronous Optical Network (SONET) standard, each ring consists of two optical fibers which form a complete circuit interconnecting a group of central offices. During normal operation, information travels in only one direction. Traffic between any pair of nodes on the ring must travel in the same clockwise or counter clockwise direction. For example, in Figure 6b, traffic from B to A might travel through node C during normal operation. If any link along the ring fails, the ring structure can guarantee full connectivity by using the spare fiber. During normal operation, traffic from A to B and traffic from B to A would travel on the same fiber, with the B to A traffic making a full circuit on the ring. If the link connecting A and B is damaged, then A to B traffic would be re-routed on the second fiber and would also make a full circuit of the ring.

SONET rings are used primarily in metropolitan area networks where the traffic volumes are moderate and the cost of redundant fiber is manageable. Networks based on ring structures, however, are difficult to manage when traffic grows at different rates throughout the network. Until recently, in the longer distance intercity network, the cost of redundant capacity has been a more

[34] In this configuration, the transmission path is not direct, but capacity is allocated (or dedicated) to specific office pairs. Circuits are converted from optical to electrical signals but are not switched at intermediate offices. Current wave division multiplexing technologies further eliminate the need for optical to electrical conversion at intermediate locations.



(a) Mesh-like Interoffice Network



(b) Interoffice Ring Network

Fig. 6. Two possible interoffice networks.

significant factor. Reliability generally has been achieved by carefully designed alternative routing algorithms based on a mesh like architecture[35].

As the volume of data traffic on the telecommunications network grows, including internet backbone traffic, the need for optical to electrical conversion is becoming a significant capacity constraint on long haul portions of the network. The current generation of switches and routers is based on digital electrical switching which requires a conversion of light waves to electrical pulses at every switching node, even if only a small portion of the signal terminates at that location. All optical networks would allow optical signals to bypass the optical/

[35] Wright (2000, p. 2).

electrical conversion by relying on new technologies using microscopic mirrors or other optical switching devices[36]. As these technologies mature, they can be expected to profoundly affect the design of future networks.

3.5. Interconnection arrangements and costs

This section considers in more detail the nature of the costs that telecommunications firms incur during interconnection. From the point of view of an individual network, an interconnecting network should be viewed as another customer. Indeed it may or may not be possible to distinguish between a network customer and a business or residential end user customer. Interconnection costs can be divided into three categories: (1) costs incurred in provisioning each network for the seamless transfers of calls through appropriate software modifications of switches; (2) the cost of dedicated transport facilities connecting two networks; and (3) the cost of switching and interoffice trunking capacity on each network to accommodate the additional peak traffic generated by the interconnection of the networks.

Costs not directly attributable to interconnected networks: Both wireline and mobile carriers incur costs in connecting end-users to their networks. Non-traffic sensitive costs attributable to subscribers of a wireline local network should, in principle, be recovered by means of a flat rate charge to the customer causing the cost. That is, the price that an individual pays for access to the network should be set equal to the incremental cost of connecting that individual to the network. Customers of wireless mobile networks incur traffic-sensitive costs in connecting to the nearest local switch, since airtime is shared among multiple users. These costs might, in principle, be recovered in an efficient interconnection pricing regime[37].

Networks may also incur non-traffic sensitive costs that cannot be attributed to any particular subscriber. These costs are not traffic sensitive, nor are they directly attributable to any particular interconnecting network. Networks also incur overhead costs for corporate operations and other network functions (e.g. for universal service and emergency services).

Non-traffic-sensitive costs of facilities connecting two networks: The dedicated facilities connecting two local networks, like end users' loops and the dedicated entrance facilities that long-distance carriers use to interconnect with local switched networks are generally non-traffic sensitive. The costs of these facilities are commonly shared and jointly provided on a 'meet point' basis, with neither carrier paying the other for the portion of the facility that the other provided. Alternatively a network provider entering a market where a network provider already

[36] Iler (2000).
[37] See the chapter by Mark Armstrong in this *Handbook*.

has operations could pay the incumbent network a non-traffic sensitive charge for the use of the incumbent's facilities.

Traffic-sensitive network capacity costs: Traffic sensitive costs are generally a function of the volume of traffic during peak load periods, not total traffic, since peak period demand determines how much capacity a network must construct and the cost of carrying an additional call off-peak is virtually zero. Cost based interconnection charges would recover these costs through charges for network capacity, rather than through uniform charges based on aggregate minutes of use.

3.6. Broadband standards and technologies

Previous sections have been primarily concerned with the design of a narrowband voice grade network. Most of the same general principles discussed there apply also to broadband networks. In the local loop, business users typically demand high capacity links on T1 or T3 connections for a mixture of voice and data traffic, and the network design principles described previously are able to account for these demands. For residential demands, the subscribers generally gain access to the network by means of copper distribution plant, which has limited bandwidth. Higher bandwidth services can be provided on copper plant for customers located sufficiently close to the switch or to a remote terminal using xDSL technologies. Cable television networks, which use a combination of coaxial copper distribution and fiber feeder plant, can also be used to deliver high bandwidth services to residential subscribers. Since the cable access technology is a shared resource, the design principles for cable are different than those for traditional wire line telephone plant. Wireless carriers are expected to begin to deliver higher bandwidth services in the near future, but a discussion of the costs and characteristics of these technologies will not be considered further in this section.

Broadband telecommunications services, including high speed data and video entertainment services, have been made possible by the rapid technological advances that have occurred in both transmission and switching in recent years. The transition of the network from analog to digital technologies has made it possible for these newer high bandwidth data services to share capacity on networks that also carry the voice traffic on the public switched telephone network. Currently, the volume of data traffic is approximately equal to the volume of voice traffic, but the proportion of data traffic is expected to grow rapidly in the future.

In order to accommodate the projected growth of new data services, telecommunications networks are evolving from 'connection oriented' to a 'connectionless' technologies. In a connection oriented or circuit switched service, a path must be established through the network before any data is transferred. Call setup procedures, therefore, constitute a significant fixed cost when the

volume of data transmitted is small. In a connectionless service, there is no overall network management or control. Individual packets are routed toward their destination on the best available facility at the time the packet is sent. Subsequent routing decisions must be made at each switching center in the network, and consequently the network intelligence must be distributed throughout the network so as to make this capability feasible.

The technical challenge in the design of a broadband network is to create standards that allow for a mix of services incorporating different transmission characteristics. One standard, known as Asynchronous Transfer Mode (ATM), was selected by the International Consultative Committee on Telephone and Telegraph (CCITT) as the basis for broadband ISDN in 1988[38]. Since then, a number of TCP/IP based protocols, such as INT-SERV (providing quality of service guarantees) and DIFF-SERV (providing scalable architectures for varying traffic intensities), as well as MPLS (Multiprotocol Label Switching) have been developed to allow for the support of integrated service networks[39].

Table 4 represents the characteristics of four classes of services that the Asynchronous Transfer Mode (ATM) technology was designed to accommodate. The last row in Table 4 refers to the need to coordinate the timing in which packets are reassembled at the receiving end with the timing with which packets are sent. Voice and video applications can both accept some amount of end to end delay, but are sensitive to variability in that delay. Excessive delay variability, called 'jitter,' is not acceptable in real time delivery of voice and video services.

4. Cost proxy models of the telecommunications network

Engineering process (cost proxy) models have been developed in recent years as an alternative to more traditional econometric and accounting approaches to estimate network costs. Because econometric models rely on assumptions of (smooth) functional forms, engineering process models offer a more detailed view of cost structures than is possible using econometric data. In addition, engineering models are better suited for modeling forward-looking (long run) costs since they rely much less on historical data than econometric models. In the telecommunications industry, cost proxy models were developed initially for various regulatory purposes, but as these models have advanced in sophistication, it has

[38] Wright (1993, p. 121).

[39] See Braden, et al. (1994), Blake et al. (1998) and Rosen et al. (2001). The most important Internet standards body is the Internet Engineering Task Force (IETF). Membership in the IETF is open, and technical work of the IETF is done in working groups organized by topic (e.g. routing, transport, security, etc.). See Internet Engineering Task Force, *Overview of the IETFG*, http://www.ietf.org/overview.hmtl.

Table 4
Classes of service supported by ATM format

Voice, Video, DS1, DS3, ISDN-B	Variable Bit Rate Voice and Video	X-25, ISDN-D, Signaling, Frame Relay	SMDS, LAN Interconnect
Connection Oriented	Connection Oriented	Connection Oriented	Connectionless
Constant Bit Rate	Variable Bit Rate	Variable Bit Rate	Variable Bit Rate
Timing Relation	Timing Relation	No Timing Relation	No Timing Relation

*Source: Wright (1993, p. 129).

become apparent that they have the potential to play a larger role in applied empirical analysis. Section 4.1 reviews the regulatory applications of proxy models. Sections 4.2 and 4.3 describe one such model is additional detail and the final section 4.4 in the chapter briefly reviews some of the empirical applications of the models.

4.1. A brief survey of cost proxy models in telecommunications

Forward-looking, ex-ante, economic computer based cost models can enable regulatory authorities to estimate the forward-looking cost of network facilities and services without having to rely on detailed cost studies, prepared by incumbent local exchange carriers, that otherwise would be necessary. In addition, a publicly available cost proxy model can be useful to regulators by providing an independent check on the accuracy of ILEC cost studies. One of the first engineering process models was developed by Mitchell (1990) for the Rand Corporation under contract with GTE of California and Pacific Bell Telephone Company. This model was a significant departure from a series of earlier econometric studies, based on historical data, which sought to estimate the degree of scale economies in the local telephone industry, both preceding and following the divestiture of AT&T. The Rand model developed simple representations of the major components of the local network: loop, switching and interoffice links, and attempted to calibrate these functions using data supplied by the client companies. One year after the Rand model was published, a more sophisticated computer based proxy model, known as the Local Exchange Cost Optimization Model (LECOM) was released by Gabel and Kennet (1991). LECOM differed significantly from the Rand model by including specific computer algorithms and network optimization routines in the design of the network. Nevertheless, by current standards, both the Rand model and LECOM are relatively crude in certain aspects of their modeling process.

The first (DOS) version of LECOM modeled a typical local exchange company serving a city consisting of three concentric rectangular regions of varying

population density: a central business district, a mixed commercial and residential district, and a residential or rural district. Serving areas in LECOM always have a uniform population distribution within them and they are always rectangular in shape. Moreover, while the size of the serving area, measured as the number of access lines, is specified as a user input, the shape of the area is by default determined by an algorithm that seeks to subdivide each of the city areas into an appropriate number of neighborhoods.

In the latter half of the 1990s, cost proxy models were submitted to a number of federal and state regulatory commissions in the United States following passage of the Telecommunications Act of 1996. In the two full years following this Act, the Federal Communications Commission (FCC) engaged in proceedings on universal service, interstate access charge reform, and local exchange competition. The local competition provisions of the Act directed the FCC to implement procedures to allow potential competitors to incumbent local exchange carriers (ILECs) the right to interconnect with other carriers' networks at rates based on cost; the right to obtain unbundled network elements at cost-based rates; and the right to obtain an ILEC's retail services at wholesale discounts in order to resell those services.

The 1996 Act also directed the FCC to reform universal service support mechanisms. In a 1997 order, the Commission adopted a forward-looking economic cost methodology to calculate support levels for non-rural carriers. Under this methodology, a forward-looking computer based cost mechanism would be used to estimate non-rural carriers' forward-looking economic cost of providing services in high-cost areas. In 1998, the Commission issued a platform order in which it selected a specific proxy model, known as the Hybrid Cost Proxy Model (HCPM). In 1999 the Commission adopted a final inputs order which adopted a set of inputs for the model along with minor modifications to the platform. The HCPM will be described in more detail in Section 4.2.

During the course of the model development and evaluation process at the FCC, several industry-sponsored models were submitted for evaluation. These include the Benchmark Cost Proxy Model (BCPM) sponsored by US West, Sprint and Bell South, and the HAI model sponsored by AT&T and MCI. The HCPM ultimately incorporated elements of both of the industry models in addition to a set of new loop design and clustering algorithms developed internally at the FCC[40].

4.1.1. Modeling loop investments

Both industry-sponsored models adopted an approach similar to LECOM for defining loop distribution areas. For example, the Benchmark Cost Model

[40] See Bush, et al. (2001).

(BCM), a precursor of the BCPM, used Census block groups (CBGs) as distribution serving areas, with the justification that CBGs on average have a household count that is reasonably comparable to local serving area line counts, and, more importantly, that CBGs represent a uniform nationwide statistical representation of population by a disinterested government body. Since the original purpose of the BCM was to estimate the relative costs of serving different regions of the country for purposes of providing universal service support for high cost areas, this latter advantage was justifiably seen as an important modeling principle. An early version of the HAI model, then known as the Hatfield Model (HM) used the same modeling approach as the BCM, although it differed significantly in its choice of recommended input assumptions.

Later versions of the BCM adopted a somewhat different modeling principle of a rival industry sponsored model: the Cost Proxy Model sponsored by Pacific Bell and developed by an independent consulting firm. The resulting Benchmark Cost Proxy Model followed a grid based approach which measured population in areas defined by one degree of longitude and one degree of latitude[41]. Serving areas were constructed in the BCPM either by forming collections of grid cells until an appropriate line count was reached, or in heavily populated regions by subdividing grid cells. In rural regions, CBGs typically include a large land mass in order to maintain a roughly constant population count throughout the country. Since the BCM and HM assumed that population was uniformly distributed throughout the entire CBG, the grid based approach offered significant modeling advantages in highly rural areas. Ultimately, however, a uniform distribution within the serving area was assumed for computational tractability.

In response to the grid-based approach of the BCPM, the HM sponsors introduced a clustering algorithm in subsequent versions of the HAI model. A clustering approach uses individual customer locations rather than grid cells as primary data points. A statistical algorithm is then used to identify natural groupings of customers which can be served by a common interface point. The HAI clustering approach and the significant customer location data preparation which was required to implement it were ultimately chosen as the most accurate of the modeling approaches by the FCC in its universal service proceeding. The particular clustering method used by the HAI model, and more importantly the loop design algorithms, that remained largely intact from earlier versions of BCM and HM, were not, however, used in the FCC approach. Instead the approach of the HCPM in both of these areas was adopted based on evidence

[41] At the equator, a resulting grid cell is approximately 3000 feet square, though the exact dimensions of cells depend on the distance from the equator. In most cases a grid cell is substantially smaller than a CBG.

that indicated that the latter approach was the most accurate available approach for modeling both urban and rural areas[42].

An advantage of the clustering approach over earlier approaches is the ability of the model to explicitly maintain a maximum copper distance constraint. Satisfactory quality for both voice grade and digital services requires that copper loop plant must be restricted to a maximum distance, where the distance depends on the quality of service offered[43]. In the HCPM, once the clustering process is complete the customer locations are assigned to a set of grid cells that overlay each cluster. The grid cells resemble the grids used by the BCPM except that the cells are an order of magnitude smaller in size, 360 feet on a side instead of approximately 3,000 feet, and the population of each cell is determined endogenously by the model from the more detailed customer location data. A loop design module is then called upon to build both distribution and feeder plant to each grid cell. The smaller size of grid cells allows both distribution and feeder plant to be built to essentially the exact customer locations in the case of distribution plant, and to the exact locations of serving area interface (SAI) points in the case of feeder plant.

Both the distribution and feeder portions of the local network are designed based on minimum cost spanning tree algorithms. These seek to determine the minimum cost configuration of network links that connect customers to the SAI and SAIs to the central office switch respectively. Earlier cost proxy models; including LECOM, BCPM and HAI used different variations of a 'pine tree' routing algorithm in which locations were connected to a central point (SAI or switch) using a pre-determined set of routes along vertical and horizontal axes, with no attempt to minimize either the distance of cost of the resulting network[44].

While the loop design portions of the HCPM represent a significant advance over earlier modeling approaches, the HCPM and other proxy models generally have been criticized for failing to take account of geographical barriers, such as

[42] 'The clustering algorithm used by HAI was rejected primarily because it was done in a pre-processing stage of the model, which raised concerns about the openness and accessibility of the data used. The HCPM loop design approach was chosen on the basis of some statistical analysis by FCC staff demonstrating that the HAI approach systematically underestimated the cost of providing service in very low density areas and systematically overestimated the cost of providing service in high density areas.

[43] In the universal service adaptation of HCPM, a copper distance limit of 18,000 feet was assumed as a default based on evidence that this amount of copper plant was compatible with traditional voice grade service and an acceptable quality of digital service via modem. Alternative service qualities, such as the provision of ADSL services using copper plant would require significantly shorter copper distance.

[44] A minimum distance spanning tree algorithm minimizes the distance of the overall network. Since the costs of a telecommunications network depend both on distance and the number of access lines carried on each link in the network (which affects cable and electronics costs), the HCPM algorithms take account of both distance and line related costs.

bodies of water or extremely mountainous terrain, which real operating telephone companies are forced to recognize when they build loop plant[45]. It is likely that future modeling efforts will seek to incorporate local cost conditions in an increasingly sophisticated manner.

Most recent loop modeling efforts rely on geocoded customer location and demand information coupled with facility routing constrained to follow the existing road grid. These enhancements significantly improve the demand model. Similarly, next generation engineering algorithms will incorporate recent DLC improvements.

4.1.2. Modeling switching and interoffice investments

In terms of modeling switching and interoffice investments, there have been relatively few advances in the current generation of cost proxy models over the approaches taken by Mitchell and in LECOM. All of the approaches define a simple switching cost function of the form

$$\begin{aligned} Cost = {} & GettingStartedCost + a_1^* \, AccessLines + a_2^* \, Trunks \\ & + a_3^* CallAttempts + a_4^* Minutes \end{aligned} \tag{1}$$

The difficulty in obtaining an accurate forward-looking representation of this cost has traditionally been in obtaining appropriate data to calibrate this function. Neither switch vendors nor switch buyers have been willing to publicly reveal the details of actual contracts, since the market for switching capacity is thin, and both sides have incentives to withhold the relevant information from competitors.

Interoffice investments have been modeled with increasing accuracy, much like investments in loop plant. The interoffice computation in LECOM assumes a full mesh network for interoffice communications where the amount of traffic carried between individual switches is assumed to be inversely proportional to the distance between them. While early versions of the BCM did not attempt to model interoffice costs in any manner, choosing instead to represent interoffice investment as a fixed percentage of loop plus switching investment, more recent versions of the BCPM and HAI models have incorporated algorithms which compute the SONET ring technologies which are increasingly used to link both host-remote configurations and host to tandem configurations. An international version of the HCPM also computes the cost of SONET rings.

[45] The HCPM incorporates a feature, first introduced by the BCM model, in which terrain factors are identified with the Census block in which each customer resides, and penalty weights are assigned for cost computations associated with difficult conditions. The terrain factors include minimum and maximum slope measurements, depth to bedrock, soil properties, and water table depth.

In the international arena, a cost model of both the local access network and the intercity network was prepared for the German Regulatory Authority for Telecommunications and Posts by Wissenschaftliches Institut für Kommunikationsdienste (WIK). The WIK intercity model is a hybrid of full mesh and SONET ring connections, based on a detailed traffic demand model for projected traffic flows between switches. The Japanese Ministry of Posts and Telecommunications (MPT) has also sponsored a Long Run Incremental Cost (LRIC) model of the local and inter-city telecommunications networks in Japan.

4.2. The hybrid cost proxy model

An economic cost proxy model for estimating the cost of network elements starts with an engineering model of the physical local exchange network, and then makes a detailed set of assumptions about input prices and other factors. Such models estimate the total monthly cost for each network element. This section examines both model design and models' use of variable input factors for network investment, capital expenses, operating expenses, and common costs.

Each of the models submitted to the FCC for evaluation was based on an assumption that wire centers will be placed at the ILEC's current wire center locations. Subject to this constraint, all remaining network facilities were assumed to be provided using the most efficient technology currently in use. The constraint to use existing wire center locations is not fully consistent with a forward-looking cost methodology, and it should be recognized that over time an existing wire center model may become less representative of actual conditions faced by new entrants and incumbents. For example, after existing wire center locations were chosen by ILECs, larger capacity switches became available as fiber/Digital Loop Carrier technologies came on the market. Both of these factors have significantly altered the fundamental trade-off between switching and transmission in the design of an optimal communications network. Because of ongoing advances in technology, both facilities-based entrants and ILECs may in the future choose a much different network topology that will result in different forward-looking costs than today's network.[46] A closely related issue is whether the placement of remote switching units should be restricted to existing wire center locations, and if the models should assume that every wire center includes either a host or remote switch. Similarly, wireless technologies may in the future be capable of providing narrowband telecommunication services at a lower cost than wireline technologies.

An accurate estimate of the cost of serving a wire center or a serving area within a wire center depends on a reliable forecast of customer demand patterns within

[46]Ongoing advances in packet switching technologies will also fundamentally alter the structure of local access networks.

the area, and the number of residential and business lines. Proxy models rely on census data to determine residential demand. However, because census data generally does not report the number of business lines, model designers must use indirect methods to estimate business demand. The potential for error in estimating business and residential demand creates certain difficulties. First, as noted below, fill factors or utilization rates may be expected to vary between business and residential lines[47]. Second, loop lengths are typically shorter for business lines than for residential lines. Thus, unless the differences in costs associated with different fill factors for business and residential areas happen to offset exactly the differences in costs associated with differences in loop lengths, the cost of serving an area will depend on the ratio of business to residential lines. An understanding of the magnitude of these competing effects, however, requires an accurate estimate of the number of business and residential lines in a particular area[48].

The HAI model and the HCPM incorporate access line demand data from the Operating Data Reports, ARMIS 43-08, submitted to the FCC annually by all Tier 1 ILECs[49]. These models incorporate data on the number of: (1) residential access lines, both analog and digital; (2) business access lines, which include analog single lines and multi-line analog and digital lines, PBX trunks, Centrex trunks, hotel and motel long-distance trunks, and multi-line semi-public lines; and (3) special access lines. The number of residential lines in each Census block is computed by multiplying the number of households in a block by the ratio of total residential lines, as reported by ARMIS, to the total number of households in a study area. The number of business lines in a wire center is determined by multiplying the number of employees, as reported by Dun and Bradstreet, by the ratio of business lines to employees as determined from ARMIS data. Some refinements to this process have been made which take account of the different demands for telephone use per employee. For example, service industry demand for telephone service is most likely greater than demand in the manufacturing sector.

A significant advantage of the HCPM over earlier models is that model's ability to estimate the cost of building a telephone network to serve subscribers in their actual geographic locations, to the extent these locations are known. To the extent that the actual geographic locations of some customers are not available, surro-

[47] Fill factors or utilization rates of loop plant are the percentage of a loop plant's capacity that is used in the network. Utilization rates are necessarily less than 100 percent so that capacity is available for growth or, in the event of breakage, to avoid outages. Lower utilization rates mean that carriers deploy more unused capacity, which increases the cost of loop plant.

[48] Alternatively, the differences in fill factors and loop lengths, and thus the cost of providing service to a particular area, may depend upon the density of customers, not the type of customers, in a particular area.

[49] Tier 1 local exchange carriers are companies having annual revenues from regulated telecommunications operations of $100 million or more.

gate locations can be chosen by uniformly distributing customers along the roads within each Census block.

Once the customer locations have been determined, the HCPM employs a clustering algorithm to group customers into serving areas in an efficient manner that takes into consideration relevant engineering constraints. The cluster module can also accept Census block level data as input. In this case every block which is larger than a raster cell is broken up into smaller blocks, each no larger than a raster cell, and the population of the original block is distributed uniformly among the new blocks. The clustering task is difficult because both engineering constraints and the general pattern of customer locations must be considered. There are two main engineering constraints. First, a serving area is limited to a certain number of lines by the capacity of the SAI. Second, a serving area is limited to certain geographic dimensions by current technology since as distance increases beyond a critical value service quality is degraded.

The objective of a clustering algorithm is to create the proper number of feasible serving areas. A clustering algorithm must consider both fixed and variable costs associated with each additional serving area. A fixed cost gives a clear incentive to create a small number of large clusters, rather than a larger number of smaller-clusters. On the other hand, with fewer clusters the average distance of a customer from a central point of a cluster, and consequently the variable costs associated with cable and structures, will be larger. In moderate to high-density areas, it is not clear, *a priori*, what number of clusters will embody an optimal trade-off between these fixed and variable costs. However, in low-density rural areas, it is likely that fixed costs will be the most significant cost driver. Consequently, a clustering algorithm that generates the smallest number of clusters could be expected to perform well in rural areas.

The final step in the customer location module is to convert the above data into a form that can be utilized by the loop design module. For wire centers that are sufficiently small, it would be possible to build plant to the exact customer locations determined by the geocoded data as processed by the cluster module. For larger wire centers, which may have 20,000 to 100,000 or more individual customer locations, it would be extremely time consuming to build distribution plant directly to each individual location. As in the clustering module, an acceptable compromise between absolute accuracy and reasonable computing time can be achieved by defining a grid on top of every cluster and assigning individual customer locations to micro-grid cells within the grid. The model assumes that customers within each micro-grid cell are uniformly distributed within the cell. If multiple customer locations fall in a single micro-grid cell, the cell is divided into lots. Loop plant can, therefore, be designed specifically to reach only populated micro-grid cells, and the individual customer lots within each micro-grid. For large wire centers, the number of populated micro-grid cells will be much smaller than the number of customer locations, even when a relatively small micro-grid size is specified. Therefore, with this approach it is

possible to place an upper bound on computing time for running the model, while simultaneously placing a bound on the maximum possible error in locating any individual customer.

Given the inputs from the customer location module, the logic of the loop design module is straightforward. Within every micro-grid with non-zero population, the module assumes that customers are uniformly distributed. Each populated micro-grid is divided into a sufficient number of equal sized lots, and distribution cable is placed to connect every lot. These populated micro-grids are then connected to the nearest concentration point, called a serving area interface (SAI), by further distribution plant. During this phase of the loop design algorithms, the heterogeneity of micro-grid populations, and the locations of populated micro-grids are explicitly accounted for. Finally, the SAIs are connected to the central office by feeder cable. On every link of the feeder and distribution network, the number of copper or fiber lines and the corresponding number of cables are explicitly computed. The total cost of the loop plant is the sum of the costs incurred on every link.

Distribution consists of all outside plant between a customer location and the nearest serving area interface (SAI). Distribution plant also consists of backbone and branching cable where branching cable is closer to the customer location. Feeder consists of all outside plant connecting the central office main distribution frame, or fiber distribution frame, to each of the SAIs.

Technology choices for both feeder and distribution plant are constrained by user specified threshold values for the distance from the central office to the most distant customer served from a given terminal. The model recognizes the following distance thresholds[50].

Copper gauge crossover: the maximum distance served by 26 gauge analog copper

Copper distance threshold: the maximum distance served by 24-gauge copper

Copper-T1 crossover: the maximum feeder distance served by analog copper

T1 fiber crossover: the maximum feeder distance served by T1 technology

The model assumes the following progression of technologies as distance increases: 26-gauge copper, 24-gauge copper, T1 on copper, and fiber[51].The model treats these thresholds as constraints in a cost optimization. For example, if relative prices indicate that fiber is less expensive than T1, fiber will be chosen even if the feeder distance is less than the T1-fiber crossover. HCPM explicitly computes the cost of the feeder plant which connects each terminal to the central office, and, subject to the distance constraints described above, selects the cost minimizing technology.

[50] The first two thresholds refer to distances from the central office to the furthest customer served in a particular grid. The second two thresholds refer to distance from the central office to an SAI.

[51] That is, copper gauge crossover $\leqslant$ copper T1 crossover $\leqslant$ T1 fiber crossover.

Two algorithms are used to determine the correct amount of cable and structures that are necessary to connect each micro-grid to the nearest SAI. The first algorithm is most appropriate in densely populated clusters, in which the proportion of populated micro-grids to total micro-grids is relatively large. Backbone cables run along every other cell boundary, and connect with the distribution plant within a cell at various points. The location of the SAI divides the cluster into four quadrants. Beginning with the 'southeast' quadrant, backbone cable is run along the boundary of the first and second rows of cells, running eastward until it reaches a point directly below the SAI. Cable from cells in row one is directed upward toward the cell boundary, while cable from cells in row two feeds downward. From the cell boundary point directly below the SAI, vertical cable then completes the connection to the SAI. By continuing along every other cell boundary, but extending cable only to populated cells, it is possible to connect every customer in the quadrant to the SAI. Similar connections apply to each of the remaining quadrants.

The second algorithm generally gives a more efficient distribution network for clusters with a lower population density, where the number of populated micro-grids is smaller. In this case, the construction of an optimal distribution network within a cluster is closely related to the problem of constructing an optimal feeder network for the entire wire center, and the same algorithm is used to provide a solution. In this approach, a minimum cost spanning tree algorithm, based on Prim (1957), is used to connect each drop terminal placed within each micro-grid to a network consisting of all drop terminal locations and all SAIs.

HCPM always computes a distribution cost using the first algorithm. Since computations involving the Prim algorithm can be time consuming for large clusters, the program has been designed to 'optimize' only the lower density clusters if the user so desires. Whenever both algorithms are used, distribution cost is determined by choosing the minimum cost obtained.

The cost of copper-based T-1 terminals and fiber-based DLC (digital loop carrier) terminals are determined as follows. The terminal sizes are a function of the line capacities required. The digital signal hierarchy, typically referred to as DS-1, DS-3 and OC-n, where n is a multiple of 3, is used to size the terminal requirements. Each DS-1 or T-1 may be used to support 1–24 lines. A DS-3 or T-3 may be used to support 672 lines (28 DS-1's). Finally, an OC-3 (3 DS-3's) may be used to support 2016 lines. Five terminal sizes with line capacities of 2016, 1344, 672, 96 and 24 may be used[52]. Fiber terminals require 4 fibers per terminal. T1 terminals, with line capacities of 24 or 96 require two copper pairs per DS-1 line: one for transmitting signals from the CO to the primary SAI, and one for receiving signals from the primary SAI at the CO. A user-adjustable 4 to 1

[52] In the international implementations of HCPM, described in the following section, the DLC terminal capcities, and certain other engineering constraints, are set by user inputs to the appropriate values for the local environment.

redundancy ratio is assumed in the model, which means that 10 copper pairs are required to serve each 96-line T1 terminal.

The HCPM makes two fundamental modifications to the Prim algorithm. First, the model creates a set of potential junction points along a pair of East-West and North–South axes with origin at the wire center. The algorithm permits, but does not require, these potential junction points to be used in creating the feeder network. Junction points create additional opportunities for the sharing of structure costs, and in some circumstances they can also reduce the distance between a terminal node and the central office. The second modification of the Prim algorithm is in the rule which is used to attach new nodes to the network. Rather than minimizing the distance from an unattached node to the existing network, the algorithm minimizes the average total cost of attaching an unattached node, and of constructing all of the lines that are required to carry traffic from that node back to the central office.

In the construction of the feeder network, the HCPM allows the user to determine whether to use airline distances between nodes or rectilinear distances. The model also applies a road factor to all distance computations in the feeder network. This road factor is intended to convert distances determined by the distance function into actual route distances, which must account for the existing road network and other terrain factors. Road factors could be determined empirically for each region of the country by comparing actual feeder route distance to model distance computation.

4.3. International applications of HCPM

The HCPM is well suited for export to other countries since the model source code and executable files are freely available and the model was designed to explicitly use a wide variety of customer location data sources. In order to adapt the model to a local environment, it is necessary only to develop a source of customer location data and to adjust the input prices and engineering constraints used by the model to values appropriate for local conditions.

At this writing, three countries outside the United States are in varying stages of adopting some form of HCPM for use as a regulatory tool. In Argentina, a research team at the Universidad Argentina de la Empresa (UADE) has developed a customer location data source and recommended changes to model logic suitable for the Argentine situation. The HCPM has also been used by the Argentine Secretary of Communications in the Regulation on Interconnection passed on September 2000. In Portugal, the HCPM has been adapted by the independent quasi-governmental authority Instituto das Comunicações de Portugal (ICP). In Peru, the HCPM is currently being evaluated by the regulatory authority El Organismo Supervisor de Inversión Peruano en Telecomunicaciones (OSIPTEL). In this section, we briefly describe the implementation process in Argentina and Portugal.

In Argentina, UADE devoted the first portion of its effort to identifying two representative areas for analysis. In the Argentine model of liberalization, the country was divided into two 'LATA-like' regions in order to create two distinct franchises (which are ultimately intended to be open to completely free entry). These franchises were sold at auction to Telefónica de Argentina, a subsidiary of Telefónica de España, and to Telecom Argentina, a subsidiary of France Telecom. Two cities – Córdoba from the Telefónica franchise area and Mendoza from the Telecom's franchise area – were chosen for the initial study.

The UADE team gathered data from Argentina's Instituto Nacional de Estadistica y Censos (INDEC) at the *manzana*, or city block, level for Córdoba. The INDEC data include information on the number of residences, number of business locations, and number of employees at those locations (business data are only available at the *radio* level, which is analogous to a Census block group in the U.S., and these are distributed uniformly across manzanas). The HCPM is capable of handling aggregated customer location data. Its fundamental design is set up to handle individual customer locations, but its creators always understood that tradeoffs between data accuracy and openness and/or availability of data may exist, so the model includes means of utilizing data at almost any resolution.

Manzanas are assigned to wire centers according to their proximity to each central office. At this time, precise data on wire center boundaries are not available in Argentina. Residential line counts are assumed to equal either the number of residences or the number of residences 'trued-up' by weighting according to the total residential line count for each wire center. Business line counts are assigned by dividing total business line count for the wire center evenly over all business locations within the calculated wire center.

HCPM has also been designed to accept detailed geological and topographical data, such as maximum and minimum grade of elevation, soil type, depth to bedrock, and water table. Unfortunately, such data do not exist in Argentina as yet, so it was decided to use average values in order to run the model.

In Portugal, the ICP modeling team developed a customer location database in an innovative way. No database exists for Portugal that includes georeferenced customer locations with an indication of demand at each point. Furthermore, population data collected by the Instituto Nacional de Estatísticas reports only at the level of the *freguesia*, an administrative unit roughly equivalent to a U.S. township. The Portuguese Instituto Geográfico do Exército, or Army Geographic Institute (IGE), has developed digitized maps showing the location and shape of every building, road, and geographical feature in the country. These data have been manipulated to work with HCPM as follows.

Each shape record is assigned to a freguesia based on the coordinates of its centroid using MapInfo® software. Once all shapes within a freguesia are

known, the number of business and residential lines in that freguesia are assigned to each location proportionally to the area of the shape at that location. Residential line counts from the freguesia are derived in one of two ways, either by using IGE data on number of residences within the freguesia, or by allocating total wire center line counts to the freguesia based on an allocation rule. When the computer program developed at ICP has assigned a line count to each shape, it then assigns each shape to a wire center by using either a minimum distance criterion (the shape is attached to the nearest wire center) or by accepting information provided by the user on wire center boundaries. Finally, at the option of the user, total lines within each wire center can be 'trued-up' to reflect line counts for the wire center.

Partial data on soil type, depth to bedrock, and water table are available for portions of Portugal, and these data have also been applied to the model. The ICP team also collected altimetry data to use in calculation of slope, or grade angle. The altimetry data exist at a resolution of approximately 100 meters latticed throughout the country. A computer program developed at ICP calculates the slope at each point by measuring the change in altitude between the point and each of its neighbors, recording the maximum value. Each value is converted into a grade angle using the appropriate trigonometric identity. A database is thus created for the entire country. MapInfo® assigns each slope value to a freguesia, and maximum and minimum values for that freguesia are calculated.

A switching and interoffice model was also developed for Portugal. Data were collected on the locations of each local and tandem switch, and for each local office the number of access lines, peak period traffic and total traffic. A clustering algorithm determined whether to serve each local office by a host or a remote switch, and then created host-remote clusters which would later be interconnected by a SONET ring. The clustering process also identified clusters of host and stand alone switches to be served by each tandem switch, so that SONET rings connecting each host and stand alone switch to a tandem and interconnecting all of the tandem switches could be constructed. The path for each ring was determined using an algorithm known to give good approximate solutions to the traveling salesman problem. The resulting interoffice network is illustrated in Figure 7.

4.4. Cost proxy models as a tool for applied empirical analysis

Various methodological attempts have been made to use firm level data to estimate cost and production functions for the telecommunications industry using econometric techniques. Among the best known of these specifications is the

Fig. 7. Interoffice SONET rings in Portugal.[53]

translog cost specification popularized by Christensen, Cummings and Schoech (1983)[54]. The translog functional form approximates a wide variety of cost functions and possesses some degree of flexibility which makes it one of the most favored specifications by economists. One must recognize, however, that historically lack of satisfactory data rather than good techniques has been the problem in the vast majority of empirical studies of the production process.

The estimation of an econometric model using the translog or any other functional form requires a sufficiently large number of observations to make the estimation of all parameters practical. These data requirements can be met either through a time series on one firm, a cross-section of many firms, or a panel data set. However, all of these data types are subject to serious limitations from the perspective of policy analysis.

A time series of one firm requires retrospective data. Typically, the interval available for firms in the U.S. is at best quarterly. Thus, variation in the data comes about from examining relatively long series, which will include periods from different technological eras corresponding to data on production processes with different technological characteristics.

[53] The ring paths illustrated in Figure 7 are for illustrative purposes only and do not represent actual outputs of the HCPM interoffice model.

[54] See the chapter by Fuss and Waverman in this *Handbook* for a review of this literature.

Cross sectional data introduce a different policy problem. Firms may vary in their rate of implementation of technological innovations, which implies that even a properly specified cost function will reveal parameter estimates for firms of average efficiency. If rapid technological innovation changes the very structure of the cost function, thereby inducing the entry of new firms, then specification error may be introduced. Even if properly specified, it can be argued that basing policy decisions on industry average cost parameters may introduce unforeseen arbitrage opportunities and attendant inefficiencies.

A third approach involves the use of a cost proxy model to generate data about a firm's technology that would traditionally be estimated by econometric methods. In a series of recent papers (Gasmi et al., 1997, 1999, 2000, 2001) and a forthcoming book (Gasmi et al., 2002) the cost proxy model LECOM is used to model telecommunications cost functions in both regulated and competitive situations. This line of research stresses the role of asymmetric information in the analysis of the regulator-regulated firm relationship. An important consequence of this asymmetry is that the regulator must give up a rent to the firm, which has superior information in order to provide the firm with social welfare enhancing incentives to minimize costs. This is the fundamental rent-efficiency trade-off that regulators need to deal with when regulating public utilities.

Modern regulatory economics has relied upon incentive theory to interpret the regulator-regulated firm relationship as a revelation mechanism. This mechanism specifies, for each announced technological parameter referred to as the firm type, a production level and a cost to be realized as well as an associated compensatory payment to the firm. A critical constraint, the incentive compatibility constraint, insures truthful announcement by the firm. A further constraint, the individual rationality constraint, insures the feasibility of the relationship, i.e. that the firm agrees to participate. If aggregate costs of the regulated firm are observable *ex post*, for example through auditing, optimal regulation can be viewed as a mechanism that specifies, for each firm type, the optimal price and effort level and the associated compensatory payment or transfer to the firm. These variables, in turn, determine the informational rent that must be left to the firm. If costs are not observable, a somewhat different mechanism, analyzed by Laffont and Tirole (1993), yields similar results.

While a number of significant theoretical insights are revealed by the theoretical analysis of incentive regulation, the usefulness of this approach is ultimately derived from its ability to make empirically testable predictions. In order to quantitatively measure the social cost of the informational rents in a regulatory setting, a detailed knowledge of the firm's cost function is required. That is, while the incentive approach to regulation acknowledges that some aspects of the regulated firm's technology are unknown to the regulator, a full solution to the regulator's problem requires knowledge of the structure of that technology. The approach of Gasmi, Kennet, Laffont and Sharkey has been to use the cost proxy model LECOM to reveal the properties of a generic telecommunications

cost function. Through appropriate simulations, cost data were generated in a variety of telecommunications environments. The resulting pseudo-data were then used to estimate a translog functional form to give a smooth representation of the telecommunications cost function of the form $C = C(\beta, e, q)$, where q represents usage of local telecommunications service, e represents the effort level of the regulated firm, and β represents the technological uncertainty of the regulator.

In Gasmi et al. (1997) this approach was used to characterize certain properties of optimal incentive regulation. In a subsequent paper (Gasmi et al., 1999) the performance of different regulatory regimes were explicitly evaluated. Incentive regulation was shown to generate substantial gains relative to traditional cost plus regulation, but the efficiency gains were shown to accrue primarily to the regulated firm in the case of a pure price cap regulatory mechanism. Under a fully optimal mechanism both firms and consumers tend to benefit, and surprisingly price cap regulation with profit sharing was shown to closely approximate the optimal mechanism.

In Gasmi et al. (2000) LECOM was used to compare several different methods for financing a universal service program in a telecommunications setting with both high cost and low cost customers. Traditionally, the universal service subsidies have been funded by a deliberate policy of setting prices based on the average cost of serving both high cost and low cost customers. This form of deliberate cross subsidy is not sustainable under free entry. There are efficiency gains to be expected when entry is allowed, but raising funds to support a universal service program from alternative sources is costly. In evaluating the resulting trade-off, it was found that an entry policy which preserves the cross subsidy between high and low cost customers may, in some circumstances, lead to higher social welfare than a policy of free and unrestricted entry.

In a related paper (Gasmi et al., 2001) the proxy model methodology was used to reconsider the traditional test for natural monopoly. In this framework, aggregate social welfare was evaluated under both regulated monopoly and a variety of competitive scenarios. This evaluation differs significantly from the traditional cost based test for natural monopoly by explicitly accounting for the informational rents or profits earned by the regulated firm, the potential for non-cost minimizing behavior under certain forms of regulation, and the costs associated with interconnection of competitive firms.

5. Conclusion

The objective of this chapter has been to accurately represent the technology and structure of cost in the telecommunications industry. Any supply side analysis of market structure or performance in the industry must be based on a thorough understanding of these elements. While all of telecommunications was once

regarded as a natural monopoly, most markets are now open to competition, though there remain significant barriers to entry. There will very likely will be pockets of local natural monopoly and a consequent need for regulatory intervention in many portions of the industry for the foreseeable future. Hence, this chapter has been written both for academic economists and government policy makers.

While the industry continues to experience rapid technological progress, the underlying structure of telecommunications costs has changed only gradually over the previous century. This chapter is organized so as to portray this structure from a variety of perspectives. Section 2 is devoted to a discussion of the design principles of telecommunications networks – principles which can be expected to endure for many years and under all but the most revolutionary forms of technological change. Section 3 describes the way in which these design principles have been implemented today through a description of the actual facilities used to provide telecommunications services using a variety of currently viable technologies. Section 4 focuses even more closely on the cost structure of the industry by describing in detail some of the engineering process (cost proxy) models that have been used, as an alternative to standard econometric analysis, to quantify both the level and structure of costs in the industry.

References

Ball, M. O., T. Magnanti, C. Monma and G. Nemhauser, 1995a, Handbook of operations research and management science, Vol. 7. Amsterdam: Elsevier.

Ball, M. O., T. Magnanti, C. Monma and G. Nemhauser, 1995b, Handbook of operations research and management science, Vol. 8. Amsterdam: Elsevier.

Bell Laboratories, 1980, Engineering and operations in the Bell system, Indianapolis: Western Electric Company, Inc.

Bellcore, 1994, BOC notes on the LEC networks: Special report SR-TSV-002275, Piscataway, NJ: Bell Communications Research, Inc.

Blake, S., D. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, 1998, An Architecture for Differentiated Services, Network Working Group, Request for Comments, 2475, http://community.roxen.com/developers/idocs/rfc/rfc1633.html.

Braden, R., D. Clark and S. Shenker, 1994, Integrated services in the Internet architecture: an overview, Network Working Group, Request for Comments, 1633, http://community.roxen.com/developers/idocs/rfc/rfc1633.html.

Brock, G. W., 2002, Historical overview, in M.E. Cave, S.K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science BV, 43–74.

Bush, C. A., M. Kennet, J. Prisbrey, W. Sharkey and V. Gupta, 2001, Computer modeling of the local telephone network, Telecommunication Systems, 18, 359–83.

Christensen, L. R., D. Cummings and P. E. Schoech, 1983, Econometric estimation of scale economies in telecommunications, in L. Courville, A. de Fontenay, and R. Dobell, eds, Economic Analysis of Telecommunications: Theory and Applications, Amsterdam: North Holland.

Economics and Technology, Inc./Hatfield Associates, Inc., 1994, The enduring local bottleneck: Monopoly power and the Local exchange carriers, Boston: Economics and Technology, Inc.

Feller, W., 1957, An introduction to probability theory and its applications, Second edition, Vol. 1. New York: Wiley.
Freeman, R. L., 1999, Fundamentals of telecommunications, New York: Wiley.
Gabel, D. and M. Kennet, 1991, Estimating the cost structure of the local telephone exchange network, Columbus, OH: National Regulatory Research Institute Report, No. 91–16.
Gasmi, F., M. Kennet, J. J. Laffont and W. W. Sharkey, 2002, Cost proxy models and telecommunications policy: A new empirical approach to regulation, Cambridge, MA: The MIT Press.
Gasmi, F., J. J. Laffont and W. W. Sharkey, 1997, Incentive regulation and the cost structure of the local telephone exchange network, Journal of Regulatory Economics, 12, 5–25.
Gasmi, F., J. J. Laffont and W. W. Sharkey, 1999, Empirical evaluation of regulatory regimes in local telecommunications markets, Journal of Economics and Management Strategy, 8, 61–93.
Gasmi, F., J. J. Laffont and W. W. Sharkey, 2000, Competition, universal service and telecommunications policy in developing countries, Information Economics and Policy, 12, 221–48.
Gasmi, F., J. J. Laffont and W. W. Sharkey, 2001, The natural monopoly test reconsidered: An engineering process-based approach to empirical analysis in telecommunications, International Journal of Industrial Organization.
Iler, D., 2000, Mirrors, bubbles lead race to pure optical switching, Communications Engineering & Design.
Kaserman, D. L. and J. W. Mayo, 2002, Competition in the long-distance market, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science BV, 509–561.
Laffont, J. J. and J. Tirole, 1993, A theory of incentives in procurement and regulation, Cambridge, MA: MIT Press.
Minoli, D., 1991, Telecommunications technology handbook, Norwood, MA: Artech House.
Mitchell, B. M., 1990, Incremental costs of telephone access and local use, RAND Report R-3909, ICTF.
Prim, R. C., 1957, Shortest connection networks and some generalizations, Bell System Technical Journal, 36, 54–64.
Rosen, E., A. Viswanathan and R. Callon, 2001, Multiprotocol Label Switching Architecture, Network Working Group, Request for Comments, 3031, http://community.roxen.com/developers/idocs/rfc/rfc1633.html.
Sharkey, W. W., 1982, The theory of natural monopoly, Cambridge: Cambridge University Press.
Wildeman, G. and F. Flood, 2000, Moving towards the L-band, Corning Telecommunications Products Division, GuideLines Magazine, Winter 2000.
Woroch, G. A., 2002, Local competition in telecommunications, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science BV, 641–716.
Wright, A., 2000, Networking on the next level, Corning Telecommunications Products Division, GuideLines Magazine, Winter 2000.
Wright, D., 1993, Broadband: Business services, technologies, and strategic impact, Boston: Artech House.

SECTION II – REGULATION

This Page Intentionally Left Blank

Chapter 7

PRICE REGULATION

DAVID E. M. SAPPINGTON*

University of Florida

Contents

* I would like to thank the editors for helpful comments, and Dawn Goodrich, Salvador Martinez and Maria Luisa Corton for excellent research assistance. I would also like to thank my colleagues, Sanford Berg and Mark Jamison, and participants in the *International Training Program on Utility Regulation and Strategy* for many helpful conversations and insights. I am especially grateful to Dennis Weisman, who has taught me a great deal about incentive regulation during our many collaborations. Much of the discussion in this chapter draws heavily from our joint work.

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

The telecommunications industry has undergone dramatic changes in recent years. New products, new services, and new technologies have been introduced. The costs and the prices of many telecommunications services have also changed substantially. And a variety of new telecommunications suppliers have begun to operate in the industry. In part in response to these changes, regulatory policy in the telecommunications industry has also changed substantially in recent years. Most notably, various alternatives to rate of return regulation have been implemented in many jurisdictions. Relative to rate of return regulation, these alternatives generally focus more on controlling the prices charged by the regulated firm than on controlling its earnings. The precise manner in which prices are controlled varies with the particular alternative that is implemented. These alternatives to rate of return regulation will be referred to as incentive regulation throughout this chapter[1].

This chapter has three primary purposes: (1) to review the variety of incentive regulation plans that have become more popular in the telecommunications industry in recent years; (2) to assess the potential advantages and disadvantages of these regulatory regimes; and (3) to examine the impact that these regimes have had on the industry.

These issues are explored as follows[2]. Section 2 describes the primary forms of incentive regulation that have been employed in telecommunications industries throughout the world. Section 3 examines the extent to which these alternatives have been implemented in practice. Section 4 reviews the primary advantages and disadvantages of incentive regulation in general. Section 5 explores the merits and drawbacks of a primary alternative to rate of return regulation – pure price cap regulation – and examines how it is implemented in practice. Some common modifications of pure price cap regulation are considered in Section 6. Section 7 analyses the measured impact of incentive regulation on performance in telecommunications markets. Conclusions are drawn in Section 8.

Because the ensuing discussion focuses on regulatory plans that have been employed in the telecommunications in recent years, it does not offer a comprehensive review of the entire economics literature on incentive regulation[3]. The ensuing discussion also devotes little attention to the important issues of interconnection and access pricing, as these issues are the subject of another chapter in this volume (Armstrong, 2002). Every attempt has been made to provide a

[1] Sappington (1994) provides an alternative definition of incentive regulation, and stresses that rate of return regulation, like its alternatives, provides meaningful incentives to the regulated firm.

[2] The influence of Sappington and Weisman (1996a) and Bernstein et al. (1996), and thus my co-authors, on the ensuing discussion will be apparent. Their many insights are greatly appreciated.

[3] Reviews of the theoretical literature on incentive regulation can be found in Besanko and Sappington (1987), Caillaud et al. (1988), Baron (1989), Laffont and Tirole (1993 2000), Laffont (1994), and Armstrong and Sappington (2002).

comprehensive review of analyses of popular incentive regulation plans, especially price cap regulation. Despite the best of intentions, it is likely that some important relevant works have not been afforded the coverage they deserve. Fortunately, there are now a variety of sources that offer useful and distinct perspectives on incentive regulation[4]. The interested reader is encouraged to read these other sources and the references they provide.

2. Forms of incentive regulation

Many alternatives to rate of return regulation are employed in telecommunications industries throughout the world. This section describes the primary alternatives that have been employed in recent years, referring to these alternatives as *incentive regulation*.

2.1. Banded rate of return regulation

Under banded rate of return regulation, the firm is permitted to keep all of the earnings it generates, provided the earnings constitute a return on capital that is sufficiently close to a specified target rate of return. If realised earnings exceed the maximum authorised level of earnings, the difference between actual and authorised earnings is returned to customers. If realised earnings fall short of the minimum level of acceptable earnings, the firm's prices are raised sufficiently to ensure that projected earnings fall within the band of authorised earnings. Banded rate of return regulation was employed to regulate the intrastate earnings of Chesapeake and Potomac Telephone in Virginia (in the United States) in 1993.

A typical banded rate of return plan is illustrated in Figure 1. The firm's target rate of return is 12 percent under the plan. The firm retains all of the earnings it generates as long as they constitute a rate of return on capital between 11 percent and 13 percent. The firm is not permitted to retain any earnings in excess of 13 percent, and it is protected against earnings below 11 percent.

2.2. Earnings sharing regulation

Earnings sharing regulation (sometimes called sliding scale regulation or profit sharing regulation) allows for explicit sharing of realised earnings between the regulated firm and its customers. Earnings sharing regulation has been employed

[4] These sources include Joskow and Schmalensee (1986), Sappington and Stiglitz (1987), Vickers and Yarrow (1988), Acton and Vogelsang (1989), Beesley and Littlechild (1989), Hillman and Braeutigam (1989), Mitchell and Vogelsang (1991), Train (1991), Armstrong et al. (1994), Blackmon (1994), Brock (1994), Laffont (1994), Crandall and Waverman (1995), Mansell and Church (1995), Crew and Kleindorfer (1996a), Newbery (1999), and Wolfstetter (1999).

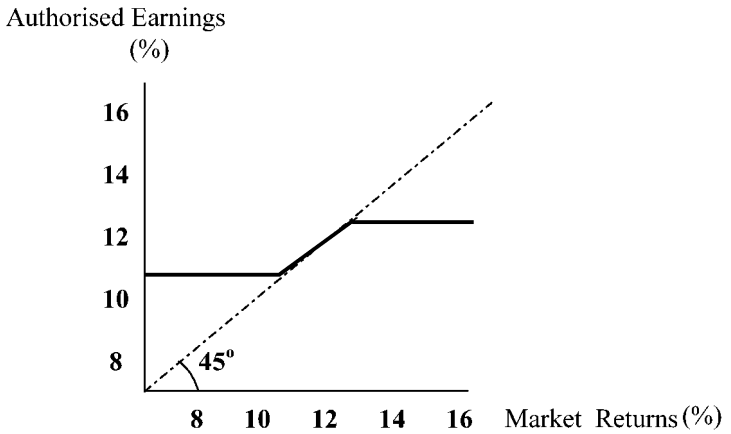


Fig. 1. Banded rate of return regulation.

throughout much of the 1990s to govern intrastate telecommunications earnings in California and New Jersey in the United States, for example. The U.S. Federal Commission has also employed earnings sharing regulation to control the earnings that local exchange carriers derive from providing inter-exchange (long distance) carriers with access to local networks.

A typical earnings sharing plan is illustrated in Figure 2. The target rate of return is 12 percent. The firm is authorised to keep all earnings that constitute a rate of return between 10 percent and 14 percent. The firm retains half of all incremental earnings between 14 percent and 16 percent. The other half of these incremental earnings are awarded to the firm's customers, usually in the form of direct cash payments or lower prices. The firm is not permitted to retain any earnings that constitute a rate of return in excess of 16 percent. Any such earnings are awarded entirely to the firm's customers.

Just as the firm shares incremental earnings between 14 and 16 percent with customers under the plan in Figure 2, it also shares with customers reductions in profit when realised earnings constitute a rate of return between 8 and 10 percent. As realised earnings decline by one dollar in this range, the firm loses only fifty cents. The other fifty cents is paid by customers, usually in the form of higher prices. If realised earnings fall below 8 percent, customers bear the full brunt of the shortfall, as prices are increased to restore earnings to a level that generates a return of at least 8 percent.

A comparison of Figures 1 and 2 reveals that earnings sharing regulation is similar to banded rate of return of regulation for realised returns that are close to or far from the target rate of return. For intermediate returns, the firm must share

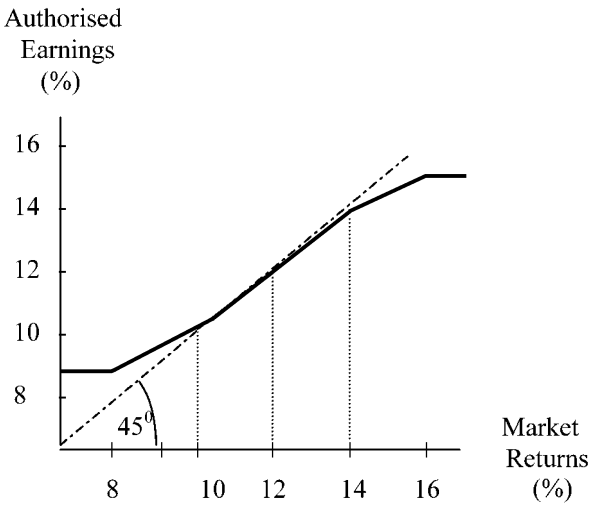


Fig. 2. Earnings sharing regulation.

incremental realised earnings with its customers under earnings sharing regulation. This sharing is the distinguishing feature of earnings sharing regulation.

2.3 Revenue sharing regulation

Revenue sharing regulation requires the firm to share with its customer's revenues (not earnings) that exceed a specified threshold. Revenue sharing regulation was implemented in the telecommunications industry in the state of Oregon between 1992 and 1996. A typical revenue sharing plan is illustrated in Figure 3. The plan allows the firm to keep all of the revenues it generates, as long as total revenue does not exceed $50 million. Once revenue exceeds the $50 million threshold, the firm is required to share with its customers half of the incremental revenues that it generates.

2.4. Rate case moratoria

Rate case moratoria are essentially agreements to suspend investigations of the regulated firm's earnings and any associated restructuring of prices to return the firm's projected earnings to target levels. The length of a rate case moratorium is usually specified in advance, and is typically in the range of two to five years. Rate case moratoria have been imposed in the telecommunications industry in many states in the U.S., including Kansas and Vermont.

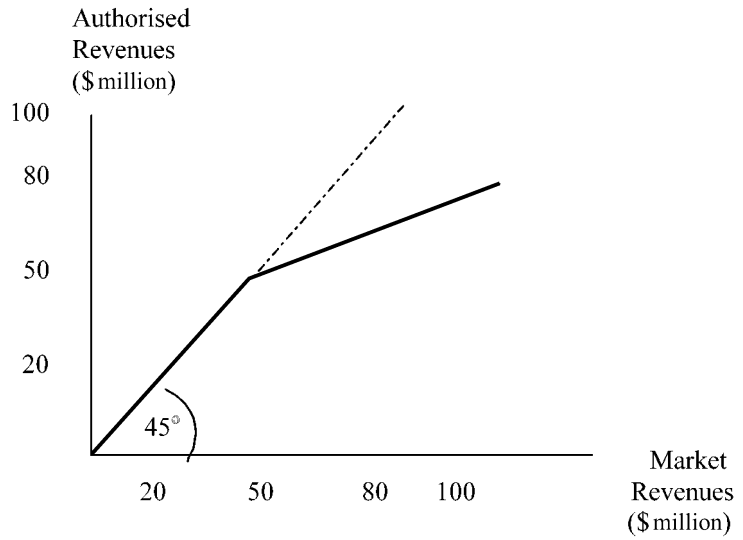


Fig. 3. Revenue sharing regulation.

2.5. Price cap regulation

Price cap regulation places limits on the prices that a regulated firm can charge, but, at least in principle, does not link these limits directly to the firm's realised earnings. Thus, in comparison with banded rate of return regulation and earnings sharing regulation, regulatory control is focused more on prices than on earnings under price cap regulation[5]. Price cap regulation generally goes beyond a rate case moratorium by specifying authorised changes in regulated prices over time. A typical price cap plan will allow the regulated firm to increase its prices, on average, at the rate of inflation, less an offset, called the *X factor*. In principle, the *X* factor should reflect the extent to which the regulated industry is deemed capable of achieving more rapid productivity growth than is the rest of the economy[6]. Price cap regulation is currently employed by Oftel in Great Britain, by the CRTC in Canada[7] and by a majority of the 50 state regulatory commissions in the United States. Price cap regulation has also been employed at various times in recent years

[5] As indicated below, the stringency of stipulated price regulations is often influenced by the firm's realised earnings in practice. In this sense, the price cap regimes that are observed in practice are seldom 'pure' price cap regimes.

[6] See Kwoka (1991; 1993b), Christensen et al. (1994), Tardiff and Taylor (1996), Bernstein and Sappington (1999a), and Section 5 below for additional discussions of this issue. Tardiff and Taylor (1996) report that the typical *X* factor is the U.S. telecommunications industry is roughly 3 percent.

[7] Oftel is Great Britain's Office of Telecommunications. CRTC is the Canadian Radio-Television and Telecommunications Commission.

in Belgium, Bolivia, France, Germany, Honduras, Hong Kong, Ireland, Italy, Japan, Mexico, Panama, The Netherlands, and Peru (OECD, 1999).

2.6. Partial deregulation

Partial deregulation is generally accomplished by classifying the services that a firm provides into different categories (e.g., basic, discretionary, and competitive), and removing regulatory controls on those services deemed to be 'competitive'. Thus, partial deregulation here entails the removal of virtually all regulatory control from some services, rather than the removal of some regulatory control from most or all services. Partial deregulation is generally implemented when competitive pressures are thought to be sufficient to keep the prices and service quality of some services at reasonable levels without direct regulatory controls.

In Great Britain, Oftel divided the interconnection services that British Telecom provides to other suppliers of telecommunications services into four categories: non-competitive services; prospectively competitive services; new services; and competitive services. The prices of non-competitive interconnection services were subject to a more stringent form of price cap regulation (i.e., one with an X factor of 8 percent) than were the prices of prospectively competitive interconnection services (which are subject to an X factor of 0 percent). New interconnection services were not immediately subject to price controls, but Oftel reserved the right to impose controls. The prices of competitive interconnection services were not regulated (Oftel, 2000)[8].

The criteria employed to distinguish between competitive and non-competitive services are not always specified clearly in practice. However, the U.S. Federal Communications Commission (FCC) has stated sufficient conditions for certain services to be removed from price cap regulation. Regulatory relief is granted for most dedicated transport and special access services, for example, when competitors have collocated and use competing transport facilities in half of the wire centres in a metropolitan area served by a regulated local exchange carrier (FCC, 1999). The FCC also permits each local exchange carrier to remove interstate,

[8] As this discussion of Oftel's regulation of interconnection services suggests, the primary benefits and costs of applying price cap regulation to retail services persist when applying price cap regulation to wholesale services. This observation has led some authors (Laffont and Tirole, 1996) to recommend global price cap regulation, under which all of the regulated firm's services (including both retail and wholesale services) are placed in a single basket and controlled with a single aggregate constraint on prices. Laffont and Tirole shown that if the regulator knows the vector of outputs that the regulated firm would produce if prices for all of its services were set at their Ramsey levels, the regulator can employ these outputs to weight permissible price changes in a global price cap regime. Doing so will induce the firm to employ its knowledge of demand and cost functions to implement prices that converge to their Ramsey levels in a stationary environment. Some complications that can arise under global price cap regulation are discussed in Sections 5.6.2 and 5.7.

intraLATA toll services from price cap regulation once it has implemented full intra- and interLATA toll dialing parity[9].

Particularly widespread deregulation of telecommunications services was implemented in the state of Nebraska in the United States. Legislative Bill 835 effectively deregulated all telecommunications services in Nebraska in 1987[10]. Only basic local service rates remain subject to direct regulatory oversight, and even this oversight is mild. The Nebraska Public Service Commission will only investigate proposed rate increases for basic local service if these increases exceed 10 percent in any year or if more than 2 percent of the telephone company's customers sign a formal petition requesting regulatory intervention[11].

2.7. Yardstick regulation

Under yardstick regulation, the financial rewards that accrue to a regulated firm are based upon its performance relative to the performance of other firms[12]. Typically, a firm that outperforms its comparison group on specified dimensions is rewarded, whereas the firm may be penalised if its performance is inferior to the performance of the comparison group. For example, a firm may receive a financial reward if the fraction of its customers that register complaints about the service quality they receive is smaller than the corresponding fraction for firms in the comparison group.

If different firms face operating environments that are inherently different, then yardstick regulation can unduly advantage some firms and disadvantage others. Consequently, yardstick regulation is not common in the telecommunications industry[13]. However, two forms of yardstick regulation have been employed[14]. First, price cap regulation can embody yardstick regulation. Recall that under a

[9] LATAs are geographic areas called local access and transport areas. Interexchange (long distance) carriers are permitted to provide telecommunications services that cross LATA boundaries (i.e. interLATA services), but the major local exchange carriers (the regional Bell Operating Companies, or RBOCs) are generally prohibited from doing so as of the time of this writing. Full intra – and inter LATA toll dialling parity is provided when customers are free to designate their preferred carrier and have all of their toll calls carried by this carrier automatically, simply by dialling the number of the parity they wish to reach.

[10] LB835, 89th Legis., 2nd session., 1986. Codified as sections 86-801 to 86-811, 75-109, 75-604, and 75-609, Reissue Statutes of Nebraska, 1943 (Reissue 1987).

[11] See Meuller (1993) for additional details.

[12] See Shleifer (1985) and Sobel (1999), for example.

[13] See Sawkins (1995), Cubbin and Tzanidakis (1998), and Diewert and Nakamura (1999), for example, for analyses of yardstick regulation in other industries.

[14] A third form of regulation that might be viewed as yardstick regulation has been employed in Chile since 1987. The benchmark to which Teléphonos de Chile is compared is the performance of a hypothetical efficient firm facing the same operating conditions that Teléphonos de Chile faces. Prices are set every five years to reflect long run incremental costs of the hypothetical efficient firm, while following standard rate of return principles to eliminate extra-normal profit. The more efficient are its operations, the greater is the profit that Teléphonos de Chile secures under this form of regulation (Galal, 1996).

popular form of price cap regulation, prices are permitted to rise at the rate of inflation, less an offset called the X factor. When the X factor reflects the projected industry productivity growth rate (relative to the corresponding economy-wide growth rate), an individual firm will fare well financially when its productivity growth rate exceeds the industry average, whereas it will suffer financially when its productivity growth rate falls short of the industry average. In this manner, price cap regulation can constitute a form of yardstick regulation.

Yardstick regulation has also been employed in New York State to determine the duration of the incentive regulation plan that was implemented in 1995. The plan had an initial expiration date of 31 December 1999. However, the New York Telephone Company was afforded the option of extending the plan for two additional years if it met two conditions: (1) its realised service quality exceeded specified thresholds; and (2) its service prices were at least 4.5 percent below the prices set by other major telecommunications suppliers, as reflected in the Telephone Communications Producer Price Index (TCPPI)[15]. This second condition is a form of yardstick regulation. The reward that New York Telephone was promised for superior performance was the right to continue the prevailing incentive regulation plan. The comparison group to which New York Telephone's price performance was compared was all major telecommunications suppliers, whose prices comprise the TCPPI.

2.8. *Options*

Regulators do not always impose a single regulatory plan. Instead, they sometimes afford the regulated firm a choice among plans. For instance, the firm may be permitted to choose among different types of earnings sharing plans or between rate of return regulation and price cap regulation.

The U.S. Federal Communications Commission (FCC) provided the regional Bell Operating Companies (RBOCs) with such a choice in 1995 and 1996. Each RBOC was afforded the three options described in Table 1 and illustrated in Figure 4. Option A required a 4.0 percent reduction in inflation-adjusted access prices in return for limited earnings sharing opportunities[16]. Option B entailed a 4.7 percent reduction in these prices in return for expanded earnings sharing. Option C involved pure price cap regulation (with no earnings sharing), with a mandated 5.3 percent reduction in inflation-adjusted access prices[17].

[15] This index is calculated by the U.S. Bureau of Labour Statistics. The details of New York State's incentive regulation plan can be found in State of New York Public Service Commission (1994).

[16] Access prices are the prices that the RBOC charges to long distance service providers for connecting a call to the RBOC's network.

[17] Between 1991 and 1994, the FCC provided the RBOCs with a choice between two earnings sharing plans. Price cap regulation was not an option.

Table 1
Options in FCC's 1995 access price regulations

Option	Minimum retained earnings	50/50 Sharing of returns	Maximum retained earnings
Option A (4.0% *X* factor)	10.25	12.25–14.25	13.25
Option B (4.7% *X* factor)	10.25	12.25–20.25	16.25
Option C (5.3% *X*-factor)	None	None	None

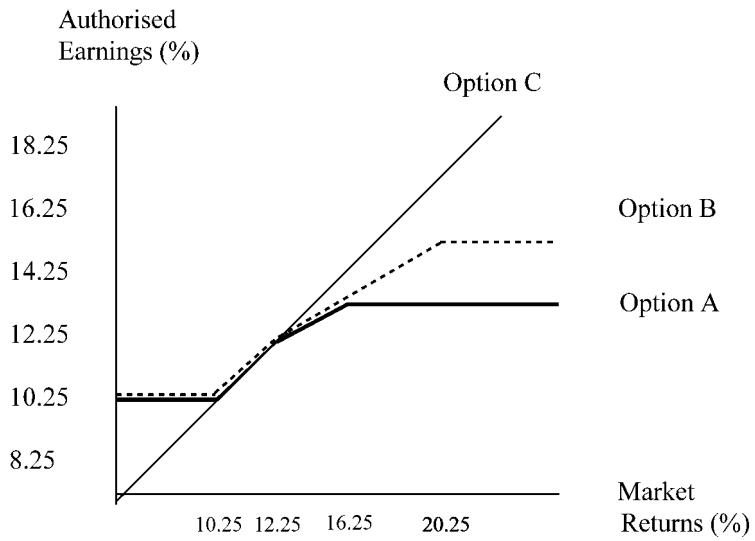


Fig. 4. Regulatory options under the FCC's 1995 plan.

Options can be particularly valuable when multiple firms with different innate capabilities are being regulated. Different firms can choose different plans when options are available. For instance, under the options presented by the FCC, an RBOC that is particularly optimistic about its ability to reduce operating costs and increase productivity can choose option C. A less optimistic RBOC can choose option A or option B. Thus, options can help to tailor regulations to the capabilities of the regulated firms. Such tailoring can secure additional gains for

consumers while limiting the likelihood of financial distress for the regulated firms[18].

Having reviewed the primary alternatives to rate of return regulation that are employed in the telecommunications industry, the discussion turns now to a review of recent trends in the implementation of incentive regulation plans.

3. Trends in incentive regulation

The purpose of this section is to examine the extent to which various forms of incentive regulation have been employed in the telecommunications industry. The most comprehensive data is available for the United States, so the discussion here will focus on the U.S. experience. The U.S. is not alone, though, in shifting from rate of return regulation to incentive regulation in its telecommunications industry. Indeed, incentive regulation was implemented in the U.S. only after price cap regulation was implemented in Great Britain in 1984[19]. And, as noted above, price cap regulation has been implemented in many countries throughout the world, including Belgium, France, Honduras, Ireland, Italy, Japan, Mexico, Panama, and The Netherlands (OECD, 1999).

Table 2 summarises the extent to which incentive regulation plans have been implemented by state regulators in the United States since 1985. (All states employed rate of return regulation prior to 1985.) The table focuses on the most popular forms of regulation[20].

Three distinct patterns are evident from Table 2. First, rate case moratoria were the most popular form of incentive regulation in the mid and late 1980's, when alternatives to rate of return regulation were first implemented in the U.S. telecommunications industry. Second, earnings sharing regulation became particularly popular in the early 1990s, as the popularity of rate case moratoria waned. Third, few states employed price cap regulation until the mid 1990s. However, by 1996, price cap regulation had become the predominant form of regulation, and it remained the predominant form of regulation at the close of the twentieth century.

The patterns exhibited in Table 2 reflect a natural progression from less radical to more radical departures from rate of return regulation. As inflation subsided

[18] Section 6.3 provides additional discussion of regulatory options, explaining how they can be employed to induce a regulated firm to employ its superior knowledge of its operating capabilities in the best interests of its customers. Also see Lewis and Sappington (1989), Sibley (1989), Sappington (1994), and Sappington and Weisman (1996a).

[19] See Armstrong et al. (1994) for a detailed review and analysis of the price cap regulation regimes imposed on British Telecom between 1984 and 1994. Also see Neri and Bernard (1994).

[20] Table 2 does not include separate categories for banded rate of return regulation or revenue sharing because of their limited use. Banded rate of return regulation is recorded as rate of return regulation in Table 2. Oregon's revenue sharing plan, which was coupled with a form of price cap regulation, is recorded as price cap regulation in Table 2. The statistics in Table 2 are derived from BellSouth Services (1987–1992), BellSouth Telecommunications (1993–1995), Kirchhoff (1994–1999), and *State Telephone Regulation Report* (2000).

Table 2
Number of states employing the identified form of regulation

Year	Rate of return regulation	Rate case moratoria	Earnings sharing regulation	Price cap regulation	Other
1985	50	0	0	0	0
1986	45	5	0	0	0
1987	36	10	3	0	1
1988	35	10	4	0	1
1989	29	10	8	0	3
1990	23	9	14	1	3
1991	19	8	19	1	3
1992	18	6	20	3	3
1993	17	5	22	3	3
1994	20	2	19	6	3
1995	18	3	17	9	3
1996	14	4	5	24	3
1997	12	4	4	28	2
1998	13	3	2	30	2
1999	11	1	1	36	2
2000	7	1	1	40	2

and major components of cost declined, regulated firms requested rate hearings to raise prices less frequently during the 1980s than they had historically. Consequently, rate case moratoria in the 1980s often served largely to institutionalise the longer time spans between rate reviews that were already occurring under rate of return regulation. When egregious profit did not arise under rate case moratoria, regulators gained the confidence required to experiment with more distinct forms of incentive regulation. Earnings sharing regulation constituted a significant, but still modest, departure from rate of return regulation. The tight bounds on allowable earnings under many earnings sharing plans helped to ensure that earnings would not depart too radically from the levels that would arise under rate of return regulation. (See Section 6.) In many cases in the late 1980s and early 1990s, realised earnings turned out to be relatively modest under earnings sharing plans. In a number of instances, earnings did not even rise to the point where the regulated firm was required to share earnings with its customers. This experience, and the knowledge gained by observing performance under less stringent forms of profit regulation, encouraged regulators to implement price cap regulation on a broad scale by the mid 1990s[21].

[21] Firms, too, may have been encouraged to solicit or embrace price cap regulation by the fact that earnings often increased under early price cap regimes.

Table 3
Patterns of state regulatory policy between 1984 and 2000

Regulatory policy	States
Persistent ROR regulation	MT, NH
ROR followed by incentive regulation, with no subsequent return to ROR regulation. (Date of switch to incentive regulation in parentheses)	AL(87), CA(90), FL(87), ID(87), IN(94), KS(90), KY(89), LA(88), MD(89), MI(90), MN(90), MS(90), NE(87), NV(91), NJ(87), ND(90), OH(95), OK(99), PA(94), RI(87), TX(91), VA(89)
ROR followed by incentive regulation, with a subsequent return to ROR regulation	AZ, AR, CT, DE, GA, IL, ME, MO, NM, NY, OR, SC, SD, VT, WA, WI

Developing competition in the telecommunications industry also enhanced the appeal of price cap regulation. Competition helped to impose discipline on incumbent providers of telecommunications services, so direct regulatory control of earnings was thought to be less crucial. Price cap regulation also served to provide incumbent providers with the expanded pricing flexibility they required to meet competitive challenges while ensuring that prices did not rise too rapidly, on average.

Despite the general transition from rate of return regulation to rate case moratoria to earnings sharing regulation to price cap regulation, this pattern was not universal. As Table 3 reveals, some states (Montana and New Hampshire) have never experimented with an alternative to rate of return regulation. Others (e.g., North Dakota in 1990 and Pennsylvania in 1994) switched directly from rate of return regulation to price cap regulation, and have not altered their mode of regulation since. Other states (e.g., Arizona, New York, and South Carolina) experimented with incentive regulation for a period of time before reverting to rate of return regulation, at least temporarily[22].

Federal regulation of the RBOCs has moved first from earnings sharing regulation to a choice between earnings sharing and price cap regulation, and then on to price cap regulation. As noted above, the FCC regulations in effect from 1991–1994 provided the RBOCs with a choice between two earnings sharing plans. In 1995 and 1996, the FCC allowed a choice among two different sharing plans and a price cap regulation plan. (Recall Table 1 and Figure 4.) In 1997, the FCC implemented price cap regulation of interstate access charges. All RBOCs were required to reduce inflation-adjusted access charges by 6.5 percent annually, and no sharing of earnings was instituted.

[22] Donald and Sappington (1995; 1997) explore some of the reasons why different states choose to implement different regulatory plans in their state telecommunications industries.

Some caution is advised in interpreting these trends and the classifications reported in Table 2. There is substantial heterogeneity among the regulatory plans within each of the categories listed in Table 2. For example, there are many different types of earnings sharing plans. Earnings sharing may be coupled with price freezes (as in California in 1995), with price cap regulation (as in Rhode Island in 1992), or with other forms of price controls. Furthermore, earnings sharing may be imposed on the regulated firm, or it may be offered to the firm as an option (as under the FCC's access price regulations in 1995 and 1996). In addition, the division of realised earnings between the regulated firm and its customers can vary across earnings sharing plans, as can the range of returns over which sharing occurs.

Even greater variation is evident among the plans classified as price cap regulation in Table 2. The amount by which inflation-adjusted prices must fall (i.e., the X factor) varies across price cap plans, as does the specified duration of the plan[23]. The amount of network modernisation or expansion that a firm must undertake in order to qualify for price cap regulation (or rate case moratorium, or other form of incentive regulation) can also vary[24]. The group of services to which price cap regulation applies can also vary across plans, as can the additional restrictions imposed on the prices charged for specific services[25]. Price cap regulation plans also vary in their treatment of the financial impact of significant, unforeseen events that are beyond the control of the regulated firm (e.g., extremely severe weather). Some plans include adjustments (called Z factors) to insulate the firm from the financial impact of severe, exogenous events, while others do not[26].

Regulatory plans in the same category in Table 2 also vary according to the service quality requirements they impose. Although some plans incorporate no explicit monetary rewards or penalties for superior or inferior service quality, other plans do. To illustrate, the price cap regulation plan imposed on New York Telephone in 1995 specifies target levels of service quality on such dimensions as the speed with which interrupted service is restored and the amount of time customers must wait to speak with a customer service representative. Failure to meet the specified targets results in substantial financial penalties under the

[23] In the British water industry, the magnitude of the X factor varies inversely with the capital investment that the regulated firm undertakes (Hunt and Lynk, 1995).

[24] To illustrate, New England Telephone was required to spend more than $280 million on network modernisation between 1989 and 1991 as a precondition for implementing a rate case moratorium in Vermont. The corresponding requirement for implementing a three-year price cap plan with earnings sharing regulation in California in 1990 was $415 million.

[25] In addition to the over-arching restriction on how rapidly prices can rise on average, many price cap plans place stringent limits on increases in basic local service rates for residential customers. It is not unusual for these rates to be frozen for the duration of the price cap plan, as they were for the seven-year plan introduced in New Jersey in 1993. See Section 5 for additional discussion of restrictions on individual prices under price cap regulation.

[26] Z factors are discussed further in Section 5.3.

New York plan, and these penalties increase over time[27]. Under the price cap regulation plan adopted in Illinois in 1995, penalties for poor service quality were imposed in the form of a higher X factor (i.e., a requirement that prices rise less rapidly). The earnings sharing plan adopted in Georgia in 1991 incorporated financial incentives for superior service quality in a different manner. Southern Bell was required to surpass specified service quality thresholds (primarily regarding the number of trouble reports per access line) in order to be eligible to share with its customers earnings in excess of those that constituted a 14 percent return on equity.

Because regulatory plans of the same type can vary substantially in detail, it is difficult to draw any general conclusions about the impact of a specified type of incentive regulation on performance in the regulated industry. However, some attempts have been made to do so. The central conclusions of these studies are reviewed in Section 7.

Before exploring the empirical findings to date, the principles that underlie the design of incentive regulation are discussed in greater detail in Sections 4 through 6.

4. General issues in the design and implementation of incentive regulation

The purpose of this section is to discuss the primary advantages and disadvantages of incentive regulation. The most common reasons for switching from rate of return regulation to some form of incentive regulation are reviewed in Section 4.1. Section 4.2 points out that many forms of incentive regulation share important features with rate of return regulation in practice. Potential drawbacks to incentive regulation are discussed in Section 4.3.

4.1. Reasons for incentive regulation

The various forms of incentive regulation described in Section 2 are often implemented in response to perceived drawbacks to rate of return regulation (RORR). The potential drawbacks to RORR include: (1) limited incentives for innovation and cost reduction; (2) over-capitalisation; (3) high costs of regulation; (4) excessive risk imposed on consumers; (5) cost shifting; (6) inappropriate levels of diversification and innovation; (7) inefficient choice of operating technology; and (8) insufficient pricing flexibility in the presence of competitive pressures. These potential drawbacks are now analysed in turn.

The defining feature of RORR is a matching of allowed revenues to realised costs. This matching limits incentives for the regulated firm to reduce operating costs. Any reduction in costs leads to a corresponding reduction in revenues, so

[27] New York Telephone was fined $46.1 million in 1996 for failure to meet specified service quality targets.

the firm and its managers perceive little gain from exerting the effort required to reduce costs toward their minimum possible levels.

By matching allowed revenues to realised costs, RORR can ensure earnings that are sufficient to guarantee investors a fair return on the capital that they provide to the firm. A fair return is the minimum amount required to convince investors to offer their capital to the regulated firm, rather than pursue alternative investments. A fair return is difficult to identify in practice. In their attempts to ensure that the regulated firm can attract sufficient capital to finance the ongoing investments required to provide high quality service to its customers, regulators may set prices to allow the firm more than a fair return on its investments. Doing so promotes over-capitalisation, as the firm expands capital-intensive investments beyond their cost minimising level, since the investments give rise to relatively large returns (Averch and Johnson, 1962). Furthermore, because the firm suffers financially when assets are removed from the rate base under RORR, the regulated firm may replace old assets with newer, more efficient assets too slowly under RORR, resulting in operating costs that are unduly high (U.S. Department of Commerce, 1991; Biglaiser and Riordan, 2001).

Since RORR can entail frequent, detailed investigations of the firm's operations and its realised costs, it can be a costly method of regulation. RORR also tends to place substantial risk on consumers. As costs change, prices also change under RORR to maintain the desired balance between revenues and costs. Consequently, in theory at least, it is consumers, not the firm, that bear the risk associated with significant changes in operating costs. In practice, consumers often bear unfavourable risk under RORR but may enjoy little of the favourable risk. When revenues fall or costs rise to levels that drive the regulated firm's actual rate of return below its authorised rate of return, the firm requests a rate hearing to raise prices. In contrast, when revenues rise or costs fall to levels that generate returns in excess of the authorised return, the firm seldom requests a hearing to lower prices (Joskow, 1974). Thus, the risk that consumers bear under RORR can be asymmetrically unfavourable.

RORR can also encourage cost shifting, inefficient technology choices, and inappropriate levels of diversification and innovation when the regulated firm operates in both regulated and unregulated markets. The incentives for cost shifting are apparent. The regulated firm gains one dollar in profit for every dollar of cost actually incurred in producing unregulated services that is counted as having been incurred in providing regulated services. Such cost shifting raises authorised revenues from regulated activities without affecting actual operating costs, thereby increasing the firm's aggregate profit.

The other distortions that RORR can invite are somewhat subtler. They arise from the matching of revenues and costs in regulated markets that RORR requires. In order to match revenues and costs in regulated markets, the firm's costs of serving its regulated customers must be calculated. Such calculations are difficult when the same inputs (e.g., facilities, personnel, and

equipment) are employed to serve customers in both regulated and unregulated markets. Joint production gives rise to common costs that are not directly attributable to operation in one particular market. For instance, when a telephone company delivers both regulated and unregulated services to a customer over the same line from the company's central office to the customer's premises, the cost of installing and maintaining the line is a common cost that is not directly attributable to either regulated or unregulated activities. But under RORR, the common costs are divided between regulated and unregulated activities, often on the basis of the relative sales of regulated and unregulated services.

Such cost allocation can invite a host of undesirable activities (Braeutigam and Panzar, 1989; Brennan, 1990; Crew and Crocker, 1991; Weisman, 1993; Brennan and Palmer, 1994). For instance, the regulated firm will tend to supply too little of its unregulated services when expanded sales of unregulated services reduce the fraction of common costs allocated to regulated activities, and thereby reduce authorised revenues from regulated services. In essence, the cost allocation procedure acts like a tax on unregulated activities, and so restricts their supply. The allocation of common costs can also provide incentives for the regulated firm to adopt other than the least-cost technology. For example, the firm may choose a technology with an inefficiently large component of fixed, common costs if those common costs serve to reduce significantly the variable costs of providing unregulated services. Similarly, if research and development costs are treated as common costs and allocated according to relative sales of regulated and unregulated services, the regulated firm can benefit financially from over-investing in projects designed to enhance profit in unregulated markets and under-investing in projects designed to improve operations in regulated markets[28].

Another drawback to RORR is that it can limit unduly the ability of an incumbent supplier to respond to competitive pressures. Under RORR, the prices of some services are intentionally set above their cost of supply while others are set below their cost of supply. For example, basic local telephone service rates for rural residential customers are set below cost in many countries in order to promote universal telephone service. Prices for many services sold to business customers in urban regions are set above cost to help offset the financial deficits incurred in providing rural residential service. Such a pricing structure provides incentives for competitors to serve the lucrative business customers and leave the unprofitable rural customers for the incumbent regulated firm to serve. When RORR requires lengthy, public hearings to revise long-standing pricing structures, the incumbent supplier is often unable to respond adequately to competitive pressures. Consequently, RORR can prevent an incumbent supplier from serving

[28] Palmer (1991) shows how diversification into unregulated markets can increase the amount of research and development conducted by a regulated firm.

some customers that it could serve at lower cost than its competitors if it had the pricing flexibility to do so.

For all these reasons, RORR has been declining in popularity in recent years. As the discussion in Section 2 reveals, price cap regulation (PCR) has replaced RORR in many jurisdictions. (Recall Table 3.) In theory at least, PCR can overcome all of these problems with RORR. By divorcing allowed revenues from realised costs, PCR provides expanded incentives for innovation and cost reduction. When the firm is permitted to retain as profit all of the reductions in costs it achieves, the firm has strong incentives to reduce its operating costs. Incentives for cost-reducing innovation are also enhanced under PCR to the extent that lower prices are induced under PCR than under RORR. The lower prices lead to higher production levels, so a given reduction in marginal cost provides a larger reduction in total cost, and thus a larger increase in profit (Cabral and Riordan, 1989; Clemenz, 1991).

PCR can also reduce technological distortions and the costs of regulation. When it does not link authorised earnings to capital or any other particular input, PCR can avoid over-capitalisation and related input distortions in production. And if reviews of price cap plans are scheduled infrequently, the costs of regulation can be reduced under PCR. Costs are reduced further when the regulated firm is authorised to change prices within well-specified bounds. By delegating pricing authority in this manner, regulators can avoid many costly and contentious hearings to analyse proposed rate changes.

PCR also shifts risk from consumers to the regulated firm. When it agrees to a price schedule that does not vary with realised costs, the firm bears all the risk – both favourable and unfavourable – associated with cost variation. When it severs the link between authorised revenues and realised costs, PCR also eliminates incentives for shifting accounting costs between regulated and unregulated activities, undertaking inappropriate levels of diversification and innovation, and adopting inefficient technologies (Braeutigam and Panzar, 1989). These undesirable incentives disappear when the firm earns an extra dollar of profit for every dollar of cost reduction it achieves, regardless of whether the reduction is realised in common costs or the costs of providing particular regulated or unregulated services. And when PCR affords the regulated firm significant freedom to vary individual service prices quickly and unilaterally, an incumbent producer will be better able to prevent less efficient competitors from serving customers[29].

4.2. *Incentive regulation in practice*

Although PCR and RORR can provide very different incentives in principle, it is not clear that they do in practice (Barnich, 1992; Waterson, 1992). The potential distinctions between PCR and RORR become blurred as they are implemented in

[29] For additional comparisons of incentive regulation and RORR, see, for example, Acton and Vogelsang (1989), Braeutigam and Panzar (1993), and Liston (1993).

practice for a variety of reasons. First, although prices may be divorced from realised costs for a period of time under PCR, the two are seldom divorced forever. When the price cap plan is reassessed at its scheduled review, ongoing price regulations are often informed by realised costs and earnings (Beesley and Littlechild, 1989). Some authors recommend that price caps be established at levels that are expected to dissipate a firm's current profit over the course of the next price cap period (Green and Rodriguez Pardina, 1999). When a price cap plan links future prices directly to realised costs and when the time between scheduled reviews of the price cap plan is relatively short, the regulated firm's incentives to reduce costs can be dulled under PCR, just as they are under RORR. Indeed, the incentives may be similar under the two regimes if the length of time between scheduled reviews of the price cap plan is similar to the time between the rate hearings that match prices to costs under RORR[30].

In some cases, prices are re-set to better match costs even before the time for the scheduled review of the price cap plan has arrived. This was the case, for example, in Great Britain in 1991. Oftel raised the X factor in the price cap plan for British Telecom from 4.5 to 6.25 that year, even though the X factor was scheduled to remain at 4.5 at least until 1992 (Armstrong et al., 1994, 27–8). Credibility problems can arise if regulators unilaterally revise the terms of specified regulatory policy before the scheduled date for reviewing the plan. If such premature intervention is expected, then no matter how strong the financial incentives for cost reduction may appear on paper, they will be seriously compromised in practice (Baron, 1991). Consequently, the potential gains from regulatory policies like price cap regulation may be minimal in settings where regulators cannot credibly promise to abide by the terms of the announced policy. In such settings, regulators are often better served by regulatory regimes (like RORR) that are more congruent with the regulators' limited commitment powers (Levy and Spiller, 1994; 1996).

Once realised costs influence allowed prices, incentives for cost shifting and inappropriate levels of diversification and innovation re-emerge. The magnitudes may be less pronounced under PCR than under RORR, but the same qualitative effects can arise under the two regimes (Weisman, 1993). In practice, price cap regimes often limit the incumbent firm's ability to change prices at will. For example, the firm is often precluded from raising politically sensitive rates (such as residential basic local service rates), even if these rates are set below production costs[31]. Price cap regulation plans can also limit the firm's ability to reduce prices, even if the price reductions serve to match the prices of competitors[32]. Thus, PCR,

[30] Some difference in incentives arises even in this case if the time between reviews is endogenous under RORR and if reviews occur only when the firm requests a review to raise prices (Pint, 1992).

[31] Giulietti and Price (2000) find little evidence that price cap regulation has caused a substantial rebalancing of prices to more closely approximate marginal costs of production.

[32] This was the case, for example, in the price cap plan that the U.S. Federal Communications Commission implemented for AT&T in 1989 (Mitchell and Vogelsang, 1991, p. 284).

like RORR, does not always afford the incumbent producer complete flexibility to respond to competitive pressures.

Although the distinctions between RORR and PCR in principle can become blurred as the regimes are implemented in practice, there is some evidence that incentive regulation is truly different from RORR, at least as it is practised in the U.S. telecommunications industry. Magura (1998) examines whether revenues are matched to costs under incentive regulation precisely as they are so matched under RORR. He finds that this is not the case between 1987 and 1994 for 34 local exchange carriers in the U.S. The closer matching of revenues to costs under RORR suggests that the forms of incentive regulation that have been implemented in the U.S. telecommunications industry may enhance incentives for innovation and cost reduction, as they permit the regulated firm to retain some of the cost savings it generates.

4.3. Possible drawbacks to incentive regulation

Even though incentive regulation may promote innovation and cost reduction both in theory and in practice, incentive regulation is not without its drawbacks. A primary drawback to price cap regulation is that it may allow prices to diverge significantly from realised production costs. The resulting allocative inefficiency can reduce aggregate welfare (i.e., the sum of consumers' surplus and profit) substantially. Furthermore, PCR can provide pronounced extra-normal profit for the regulated firm, which may have undesirable distributional consequences.

These drawbacks to PCR are most pronounced when: (1) there is considerable variation in possible costs; (2) the regulator values consumers' surplus much more highly than profit; and (3) positive production levels are always desirable, but the regulated firm can choose not to operate with impunity. When factors (1) and (3) prevail, a regulator cannot avoid the possibility that the firm will earn considerable rent under PCR. To induce the firm to operate when costs turn out to be relatively high despite the firm's best efforts to reduce costs, authorised prices cannot be too low. But relatively high prices will afford the firm considerable rent when realised costs are fortuitously low. Consequently, when factor (2) is also present, PCR may not be the best regulatory plan to implement.

Schmalensee (1989) demonstrates that earnings sharing regulation, and perhaps even RORR, can outperform PCR when factors (1)–(3) prevail. Although earnings sharing regulation and RORR limit incentives for cost reduction, they keep prices closer to realised cost and better limit the profit that accrues to the regulated firm. PCR fares better when the regulated firm need only be guaranteed non-negative *expected* profit, rather than non-negative profit for all possible cost realisations. In this case, prices can be lowered to the point where the firm's extra-normal profit when realised costs are low are offset by losses when realised costs are high. Gasmi et al. (1994) show that the ability of the firm to set prices below the maximum level authorised by the cap also enhances the performance of PCR.

Still, though, some earnings sharing is often preferable to PCR in the presence of distributional concerns and considerable uncertainty about feasible production costs.

As noted above, incentive regulation in general and PCR in particular shift risk from consumers to the regulated firm. Although this shifting of risk can help to motivate the firm to operate diligently, it can also raise the firm's cost of capital. Investors typically demand higher expected returns as the risk they are asked to bear increases. Consequently, another potential drawback to incentive regulation is that it can raise the cost of capital, thereby increasing an important component of industry operating costs[33].

A third potential drawback to incentive regulation stems from the strong incentives it can provide to reduce operating costs. One common way to reduce costs is to reduce service quality. For example, a telecommunications supplier may reduce its repair and customer assistance staffs in order to limit the wages and benefits it pays to its employees. Such staff reductions can cause service quality to decline below historic levels. If historic levels of service quality do not exceed ideal levels, then the resulting decline in service quality under incentive regulation can reduce welfare (Liston, 1993).

The pricing flexibility that is often afforded the regulated firm can constitute a fifth drawback to incentive regulation. Although pricing flexibility can enable an incumbent supplier to respond to competitive pressures and thereby prevent operation by a higher-cost rival, the flexibility can also serve to undo cross-subsidies that regulators have implemented to promote equity, fairness, and/or other political objectives. For example, regulators often set the same price for basic local telephone service across large geographic regions, even though the cost of providing the service varies greatly across the regions. In particular, the cost of providing service to an urban customer is often substantially less than the corresponding cost for a rural customer. When the regulated firm is afforded pricing flexibility in these circumstances, it will generally wish to set rates that approximate costs more closely. But by raising the rates on services that are more costly to provide and lowering the rates on services that are less costly to provide, the firm will undo the cross-subsidies that the regulator has implemented.

The regulated firm may also employ expanded pricing flexibility to deter welfare-enhancing entry into the regulated industry, particularly when average price levels are regulated (as they are under price cap regulation). When average price levels are regulated, a reduction in the price of a product for which the incumbent supplier faces intense competition authorises an increase in the price of

[33] Alexander and Irwin (1996) report that firms operating under price cap regulation experience greater idiosyncratic risk than firms operating under rate of return regulation in Canada, Japan, Sweden, the United Kingdom, and the United States. Investors generally require a higher expected return in order to invest in firms with greater idiosyncratic risk, which raises the firms' cost of capital.

a product for which the firm faces little or no competition. Consequently, the regulated firm may find it profitable to respond aggressively to competitive challenges, even to the point of pricing some products below marginal production costs (Armstrong and Vickers, 1993).

A fourth potential drawback to incentive regulation is the rate shock it can promote. Rate shock arises when regulated prices increase substantially and abruptly. Rate shock can arise under incentive regulation precisely because incentive regulation divorces prices from realised costs for a considerable period of time. If costs rise significantly during this period, and if the cost increase was largely unanticipated at the start of the price cap regime, then prices may need to be increased substantially when the incentive regulation plan is reassessed at its scheduled review date, thereby causing rate shock (Isaac, 1991).

Depending upon how they are designed and sequenced, incentive regulation plans can also provide financial incentives for strategic inter-temporal shifting of revenues and costs. To illustrate these incentives, consider a setting where PCR is followed by RORR. Since it is permitted to retain all of the profit it generates under PCR but revenues are matched with realised costs under RORR, a regulated firm in this setting may gain financially if it accelerates revenues and defers costs as the end of PCR and the beginning of RORR approaches (Isaac, 1991). Revenues can be accelerated by, for example, introducing new services relatively rapidly. Costs can be deferred by, for example, postponing routine or precautionary maintenance procedures. By accelerating revenues and deferring costs, the firm may be able to increase realised profit under PCR, without reducing the level of profit it is afforded under RORR.

Related strategic behaviour can arise even when the regulated firm operates only under PCR. To illustrate this fact, suppose that at each scheduled review of the price cap plan, realised costs in the preceding year are employed to assess likely future costs, and thus the most appropriate value for the X factor. In this setting, the firm could gain financially by shifting costs to this test year and by limiting its own cost-reducing efforts in this test year. Pint (1992) documents the substantial welfare gains that can arise if such strategic cost shifting and effort allocation under PCR is mitigated by basing forecasts of future costs on realised costs throughout an entire price cap period, rather than on costs in a particular test year[34].

These are just some of the forms of strategic behaviour that incentive regulation can promote. Additional forms of strategic behaviour are analysed in the next two sections, where some of the key considerations in the design of price cap regulation are discussed in greater detail.

[34] Vogelsang (1989) suggests a similar approach to implementing price cap regulation.

5. Designing price cap regulation

The purpose of this section is to examine in more detail some of the key considerations in the design of price cap regulation. Recall from Section 2 that price cap regulation plans generally permit the regulated firm's prices to rise, on average, at the rate of economy-wide output price inflation less an X factor. This restriction can be represented formally as:

$$\dot{p} \leqslant I - X, \tag{1}$$

where $\dot{p}$ denotes the rate of growth of the prices charged by the regulated firm, I is the economy-wide rate of output price inflation, and X is the X factor.

The discussion in this section begins by explaining (in Section 5.1) one rationale for imposing constraint (1) under price cap regulation and, most importantly, how to determine an appropriate value for the X factor. The proper duration of a price cap plan is discussed next (in Section 5.2) followed by an assessment (in Section 5.3) of the merits of allowing changes to the plan before its scheduled review date. A detailed discussion of how to implement the aggregate price constraint (constraint 1) follows (in Sections 5.4 and 5.5), before additional restrictions on individual service prices, multiple price cap constraints on distinct baskets of services, and service quality regulations are analysed (in Sections 5.6, 5.7, and 5.8).

5.1. Setting the X factor

As is evident from expression (1), the X factor imposed in a price cap plan determines the authorised growth rate of inflation-adjusted prices. Therefore, the magnitude of the X factor is a critical determinant of the level of welfare that consumers and the regulated firm achieve under price cap regulation. The higher is the X factor, the lower is the authorised growth rate of prices, and thus the higher is consumers' surplus and the lower is profit, *ceteris paribus*.

To understand the key factors that influence the proper choice of the X factor under price cap regulation (PCR), it is helpful to consider the fundamental role of PCR. Like many forms of regulation, PCR is often intended to replicate the discipline that competition would impose if it were present in the regulated industry. Competition enables firms to pass on to customers in the form of higher prices unavoidable cost increases (due to higher input prices), but compels firms to deliver to customers in the form of lower prices realised increases in productivity. (Productivity, recall, reflects the ratio of the firm's outputs to its inputs.) Therefore, in a competitive economy, prices rise at a rate equal to the difference between the rate at which input prices rise and the rate at which (total factor) productivity increases.

Now consider a regulated industry within an otherwise competitive economy. The rate of output price inflation outside of the regulated industry will reflect

the difference between the rate of input price inflation and the rate of productivity growth in the competitive sectors of the economy. Initially suppose that the regulated industry is deemed capable of achieving the same rate of productivity growth as the other sectors of the economy. Also suppose that the firms in the regulated industry face the same input price growth rates that the competitive firms face. Under these circumstances, once prices are set initially to generate zero extra-normal profit in the regulated industry, this profit level can be maintained by allowing industry output prices to rise at the rate of output price inflation elsewhere in the economy. Consequently, in this setting, competitive forces would be replicated in the regulated sector if the X factor were set equal to zero.

In this setting, the discipline of competitive markets would be replicated if the X factor reflected the extent to which the regulated industry is deemed capable of achieving more rapid productivity growth and faces a lower input price growth rate than other sectors of the economy (Kwoka, 1991: 1993b). To illustrate, suppose the regulated industry is deemed capable of achieving a 4 percent annual productivity growth rate, while the expected annual productivity growth rate elsewhere in the economy is 3 percent. Also suppose the prices of the inputs employed elsewhere in the economy are expected to increase by 1.5 percent annually, while the corresponding input price growth rate for the regulated industry is 0.5 percent. In this setting, the X factor that will provide no expected growth in extra-normal profit in the regulated industry is 2. This is the sum of the higher expected productivity growth rate $(4 - 3 = 1)$ and the lower expected input price growth rate $(1.5 - 0.5 = 1)$ in the regulated industry.

These considerations provide one set of guidelines for determining the X factor in price cap regulation plans. Although these guidelines are instructive, they do not provide all of the information required to implement price cap regulation in practice. In particular, productivity and input price growth rates can be difficult to predict. In practice, historic growth rates are often employed as the best predictors of corresponding future growth rates. Adjustments are common, though, particularly at the start of a price cap regime that follows an extended period in which rate of return regulation (RORR) was imposed on a private, profit-maximising firm, or in which the telecommunications supplier was a publicly-owned enterprise[35]. Since RORR and/or public ownership can limit incentives for innovation, cost reduction, and productivity growth, historic rates of productivity growth may understate the corresponding rates the regulated industry can reasonably achieve under PCR. Therefore, the X factor imposed in these settings is often increased beyond the level identified above by what is called a *stretch factor*. The stretch factor is an estimate of the amount by which the

[35] British Telecom operated as a public enterprise immediately before it was privatized and subjected to price cap regulation in 1984.

annual productivity growth rate in the regulated industry will exceed the historic rate because of the incentives for enhanced productivity growth provided by price cap regulation[36].

Other adjustments to the X factor warrant consideration in some settings. For example, price cap regulation is often applied to only a subset of the services supplied by a regulated firm, but the firm's historic measured productivity growth rate typically pertains to its entire operations. If competition in unregulated markets will force the firm to reduce prices more rapidly than its overall productivity growth rate will permit it to do profitably, then it can be appropriate to reduce the X factor in order to allow a higher rate of growth for regulated prices. Bernstein and Sappington (1999; 2000) characterise the appropriate adjustments. Emerging competition in the regulated industry can also affect the best value for the X factor. While increased competitive pressures can force the regulated firm to secure a higher productivity growth rate, they can also serve to lower realised productivity growth rates, particularly in the short run. As competitors attract the customers that incumbent suppliers built their networks to serve, the output growth rates for incumbent suppliers can decline more rapidly than their input growth rates decline, leading to a reduction in their realised productivity growth rates. Corrections for these varying effects of competition when setting the X factor can be important (Bernstein and Sappington, 1999; 2000).

In some countries, the data required to calculate productivity and input price growth rates might not be available. Consequently, the X factor cannot be calculated directly in the manner described above. In such settings, an appropriate X factor must be determined by other means. One possibility is to calculate historic changes in the prices of regulated services, and require the firm to implement future changes that are at least as favourable to consumers. A second possibility might be to impose an X factor that appears to have served the needs of all relevant parties well in neighbouring countries with similar characteristics. A third possibility arises in countries where similar firms serve different but comparable monopoly markets in the same industry. In such countries, the X factor imposed on each firm might be linked to the productivity growth rates achieved by the other monopoly producers in the industry[37].

Even in settings where historic productivity and input price data are readily available, regulators may deem such data to provide unreliable estimates of likely future productivity gains and input price changes. Consequently, the regulators may choose an X factor based upon explicit projections of futures revenues and costs, taking into account both current revenues and costs and anticipated changes in the industry. In particular, the X factor can be determined in much the

[36] The U.S. Federal Communications Commission included a 0.5 stretch factor in its 1989 price cap plan for AT&T. The Canadian Radio-Television and Telecommunications Commission imposed a 1.0 stretch factor on telecommunications suppliers in the price cap plan it introduced in Canada in 1997.

[37] This is a form of the yardstick regulation described in Section 2.7.

same way that an allowed rate of return is determined under RORR[38]. To the extent that the X factor is increased to reflect recent improved performance by the regulated firm, this method of calculating an X factor can limit incentives for superior performance. However, to the extent that this methodology provides more accurate estimates of potential productivity gains, it can limit the risk of affording enormous profit to the regulated firm or jeopardising its financial integrity. Alternative procedures for limiting this risk are discussed in Section 6.

5.2 Determining the length of time between reviews

The X factor is designed to present a significant but reasonable challenge for the regulated firm. If the X factor is set too low, the firm will earn substantial profit and production levels will be unduly low because prices are set far in excess of realised costs. If the X factor is established at too high a level, the firm may face financial distress. Because it is difficult to identify the ideal X factor in advance, it is wise to limit the period of time for which any specified X factor is in effect. Numerous factors influence the most appropriate length of time between reviews of a price cap plan.

One factor is the regulator's uncertainty about the environment in which he and the firm operate. When the regulator is very uncertain about likely future industry demand and cost conditions, he will find it difficult to specify an X factor that poses a significant but reasonable challenge for the firm. The difficulty is compounded when the regulator is uncertain about the firm's ability to reduce operating costs through its own diligent efforts. To reduce the risk associated with an X factor that is poorly matched to the firm's environment and capabilities, a relatively short time period between reviews of the price cap plan can be instituted (Mayer and Vickers, 1996).

Of course, depending upon the nature of the price cap review, a short time between reviews can limit the firm's incentive to reduce operating costs. Suppose the review is employed to reset prices and the X factor to levels that eliminate the firm's expected profit during the next price cap period, using any information that the firm's realised performance provides about its ability to secure low future costs. In this setting, the firm will often have little incentive to operate diligently if the time period between reviews of the price cap plan is short. This is because realised cost reductions during a price cap period will only increase profit for a short period of time, and will encourage higher X factors in subsequent periods[39]. The firm may be less reluctant to reveal its ability to secure low production costs if

[38] See Cave (2000) and Green and Rodriguez Pardina (1999) for details, including details on how current revenues and costs can be adjusted to account for forecast variation in future revenues and costs.

[39] These considerations also arise in determining the optimal regulatory lag under RORR. A longer lag encourages cost reduction, but increases the length of time for which prices exceed operating cost (Bailey and Coleman, 1971; Baumol and Klevorick, 1970; Pint, 1992).

the regulator can credibly promise not to set future X factors to eliminate all expected (extra-normal) profit for the firm, based upon its observed performance. One way to implement such a promise is to set the X factor to eliminate average expected profit based upon industry performance rather than the performance of individual firms in the regulated industry. Such a procedure serves to reward firms that secure the lowest operating costs and penalises those with the highest realised costs[40]. When a procedure of this sort is employed, a shorter lag between reviews of the price cap plan can be instituted to keep prices aligned with costs without dulling incentives for cost reduction unduly.

The ideal lag between reviews of a price cap plan will also vary with the firm's ability to reduce operating costs and with the price elasticity of demand for the firm's product. When demand is inelastic, relatively little economic surplus is sacrificed when prices exceed marginal production costs. Consequently, regulatory reviews to match prices to realised costs are less crucial, and so longer lags between reviews can be implemented to encourage innovation and cost reduction. Therefore, longer lags between reviews of the price cap plan are generally advisable when demand for the service of the regulated firm is inelastic and when the firm is believed to have considerable ability to reduce realised operating costs through its own diligent efforts (Armstrong, Rees and Vickers, 1995).

The regulator's ability to allow the regulated firm to earn extra-normal profit also influences the ideal length of time between reviews of a price cap plan. If political or ideological considerations render extremely high or extremely low profits impossible for the regulator to tolerate, then the price cap plan will have to be reviewed frequently to realign prices and costs. Of course, when a regulator cannot tolerate significant variation in realised profit, he may be better served by a regulatory plan other than price cap regulation (Levy and Spiller, 1996).

The ideal lag between reviews of any regulatory plan will also depend upon the details of the plan. For instance, if politically-sensitive rates (such as residential basic local service rates) are frozen under a price cap plan, then the regulator need not be concerned that these rates will rise unduly before the firm's pricing structure is re-evaluated at the price cap review. Consequently, a longer time between reviews can be advisable. Similarly, if an incentive regulation plan includes elements like earnings sharing that serve to limit extreme variation in profit between reviews of the regulatory plan, then frequent reviews are a less important means of preventing unacceptably large or small levels of profit. In practice, price cap plans tend to be reviewed every four or five years. Over time, as experience with price cap regulation has increased, the typical time between reviews of price cap plans has increased gradually (Kirchhoff, 1994–1999).

[40] If the X factor for a firm is set to eliminate its expected profit based upon the performance of the other firms in the industry rather than upon the industry average, then the link between a firm's current performance and the stringency of the future regulations it faces can be severed completely.

5.3. Mid-stream corrections: Z factors

Extreme variation in profit during a price cap regime can be limited by permitting adjustments to the price cap before the scheduled review of the regime. Such adjustments (referred to as Z factors) are often permitted for events that have three distinguishing features. First, the events and their financial implications are beyond the control of the regulated firm. Second, the events affect the regulated firm disproportionately. Third, the events have pronounced financial impacts.

The first feature – that the events are exogenous – is designed to avoid reimbursing the regulated firm for financial shortfalls that arise from its own mistakes. The second feature – that the events affect regulated suppliers disproportionately – avoids double counting. If an event that raises the costs of regulated suppliers also raises the costs of all other suppliers in the economy, then the inflationary impact of the event will be reflected in the inflation component of the price cap (i.e., the I term in expression (1)), so an additional adjustment is not required. The third feature – that the events have large financial implications – limits prolonged and costly regulatory hearings regarding events that are of limited economic importance.

Events that are often characterised by these three features include: (1) new taxes that are levied exclusively on regulated suppliers; and (2) unpredictable natural disasters that increase the operating costs of regulated suppliers disproportionately relative to other producers.

5.4. Implementing the price cap constraint

As noted above, price cap regulation typically serves to limit the average rate of growth of the prices charged by the regulated firm. Therefore, to implement price cap regulation, it is necessary to specify precisely how the average growth rate of the firm's prices will be calculated. The purpose of this subsection is to: (1) explain why a weighted average of the growth rates of individual prices is more appropriate than an unweighted average; (2) explore the advantages and disadvantages of different weighting procedures; and (3) examine the merits of affording pricing discretion to the regulated firm under different weighting procedures. The particular implementation of price cap regulation that is employed most often in practice is reviewed next.

5.4.1. The rationale for a weighted average

The average growth rate of a firm's prices can be calculated in different ways. For example, an unweighted average of the growth rates of individual prices might be constructed. Although this statistic would be simple to calculate, it would generally fail to reflect accurately the true impact of price changes on consumers. A given price change will have a greater effect on consumers the more of the

relevant commodity they are consuming, *ceteris paribus*. Thus, the average rate of growth of prices is best represented by a weighted average of individual growth rates, where the weight placed on each rate reflects its relative impact on consumers. Weights that are commonly employed in practice include the relative quantities and the relative revenues of the products supplied by the regulated firm.

One rationale for quantity weights is the following. Suppose the objective of price cap regulation is to ensure a specified level of consumers' surplus, S^0, over time[41]. Then, after setting prices ($p \equiv (p_1, \ldots, p_n)$) for the n regulated products initially to secure S^0, the regulator would want to ensure that any price changes implemented by the firm did not reduce consumers' surplus, S. For small price changes ($\mathrm{d}p_i$), the relevant restriction is:

$$\sum_{i=1}^{n} \frac{\partial S}{\partial p_i} \mathrm{d}p_i \geqslant 0. \tag{2}$$

Expression (2) states that the combined effect of all of the price changes implemented by the firm is to (weakly) increase consumers' surplus.

With independent demands and no income effects, a small increase in the price of the firm's ith product reduces consumers' surplus by the amount of the product purchased by consumers, so $\partial S/\partial p_i = -Q_i(p_i) \equiv -q_i$[42]. Therefore, expression (2) can be rewritten as:

$$\sum_{i=1}^{n} (-q_i)\mathrm{d}p_i \geqslant 0 \qquad \text{or} \qquad \sum_{i=1}^{n} q_i \, \mathrm{d}p_i \leqslant 0. \tag{3}$$

Expression (3) says that a set of proposed price changes will not reduce consumers' surplus if the sum of the price changes, each weighted by the relevant quantity of the product sold, is less than or equal to zero (Brennan, 1989; Vogelsang and Finsinger, 1979). Thus, the relevant weighted average of price changes is calculated using current output levels as weights.

5.4.2. *Tariff basket regulation*

When the demand functions facing the regulated firm are not known precisely (as they seldom are in practice), the amount of output consumers will purchase at any proposed set of prices cannot be predicted perfectly. Consequently, some approximation of expression (3) must be employed in place of the expression itself. One natural approximation replaces output levels at the proposed prices

[41] This might be the case if, for example, the rate of output price inflation in the economy and the relevant X factor were both zero.

[42] $S(p_1, \ldots, p_n) = \sum_{i=1}^{n} \int_{p_i}^{\infty} Q_i(\tilde{p}_i)\mathrm{d}\tilde{p}_i$, where $Q_i(p_i)$ is the demand function for the firm's ith product. Therefore, $\partial S(\cdot)/\partial p_i = -Q_i(p_i)$.

with output levels at the current prevailing prices. Thus, letting p_i^0 denote the current price of the firm's ith product and p_i^1 the corresponding proposed price, expression (3) could be approximated by:

$$\sum_{i=1}^{n} q_i^0[p_i^1 - p_i^0] \leqslant 0 \qquad \text{or} \qquad \sum_{i=1}^{n} p_i^1 q_i^0 \leqslant \sum_{i=1}^{n} p_i^0 q_i^0. \tag{4}$$

If expression (4) were to constitute the central constraint under price cap regulation, the firm would be permitted to charge prices ($p^1 \equiv (p_1^1, \ldots, p_n^1)$) that, when evaluated at prevailing output levels ($q^0 \equiv (q_1^0, \ldots, q_n^0)$), do not increase the firm's revenue above its present level ($\sum_{i=1}^{n} p_i^0 q_i^0$). Thus, a Laspeyre's revenue index is not permitted to increase under this form of price cap regulation, which is often referred to as *tariff basket regulation* (Armstrong et al., 1994).

Before reviewing the effects of tariff basket regulation, consider the rationale for affording the regulated firm any pricing discretion. The regulator could simply prohibit the firm from altering the initial prices (p^0) that generate consumers' surplus S^0. Doing so would ensure that consumers' surplus never falls below S^0. But doing so would also preclude any increase in consumers' surplus. Viewing current output levels as exogenous to the firm, constraint (4) ensures (weak) increases in both consumers' surplus and profit relative to a requirement that prices remain at p^0(Armstrong and Vickers, 1991). The increase in consumers' surplus arises for the following reason. Constraint (4) requires that any price increases be at least offset by corresponding price decreases, using prevailing output levels to weight the price increases and decreases. But price changes that satisfy this constraint actually benefit consumers. The benefit arises because consumers reduce their purchases of products whose prices have been increased and increase their purchases of products whose prices have been reduced. Thus, the actual impact on consumers of proposed price changes is more favourable than the impact that is calculated using prevailing output levels. Therefore, pricing flexibility subject to constraint (4) provides gains for consumers relative to the status quo.

Now consider the impact of imposing constraint (4) on the regulated firm each year, so that the relevant restriction it faces on the prices it changes in year t($p^t \equiv (p_1^t, \ldots, p_n^t)$) is:

$$\sum_{i=1}^{n} q_i^{t-1}[p_i^t - p_i^{t-1}] \leqslant 0 \qquad \text{or} \qquad \sum_{i=1}^{n} p_i^t q_i^{t-1} \leqslant \sum_{i=1}^{n} p_i^{t-1} q_i^{t-1}. \tag{5}$$

For the reason just explained, consumers' surplus (weakly) increases each year when tariff basket regulation of this type is imposed, provided the demand functions facing the regulated firm do not change over time. Furthermore, in a stationary environment, prices converge to levels that maximise consumers' surplus subject to providing a particular level of profit ($\tilde{\pi}$) for the firm (Brennan

1989; Vogelsang, 1989)[43]. The magnitude of this profit ($\tilde{\pi}$) depends upon the initial price levels (p^0). To illustrate, if the initial prices are profit-maximising prices, the firm will implement those prices each year, and $\tilde{\pi}$ will be the level of profit an unregulated monopolist would secure.

To better limit the firm's profit under price cap regulation, a more stringent constraint might be imposed on the firm. For example, instead of restricting prices to generate (weakly) less than the current revenue when evaluated at current output levels as in expression (5), prices might be restricted to those that, when evaluated at current output levels, reduce revenue by at least the amount of current profit. Formally, constraint (6) might be imposed in each year:

$$\sum_{i=1}^{n} p_i^t q_i^{t-1} \leqslant \sum_{i=1}^{n} p_i^{t-1} q_i^{t-1} - \pi^{t-1}, \quad \text{where } \pi^{t-1} \equiv \sum_{i=1}^{n} p_i^{t-1} q_i^{t-1} - C(q^{t-1}). \tag{6}$$

In expression (6), $C(q^{t-1})$ denotes the regulated firm's total cost of producing output vector $q^{t-1} \equiv (q_1^{t-1}, \ldots, q_n^{t-1})$.

When constraint (6) is imposed each year in an environment where demand and cost functions do not change over time, consumers' surplus increases each year and the Ramsey optimum is ultimately secured, provided the firm acts to maximise profit each year (Vogelsang and Finsinger, 1979). At the Ramsey optimum, the highest possible level of consumers' surplus is achieved subject to ensuring zero (extra-normal) profit for the firm (Ramsey, 1927)[44].

Despite this attractive feature of constraint (6), the constraint is not without its potential drawbacks. The constraint links allowed prices to realised profit and thus to realised cost. Consequently, the firm's incentive to minimise production costs may be dulled. Under some conditions, the regulated firm can gain financially from wasting resources in order to inflate its realised costs when constraint (6) is imposed. The higher costs reduce measured profit, and thereby relax constraint (6). The induced increase in costs can be so severe as to cause aggregate

[43] Neu (1993) illustrates some of the distortions that can emerge when demand functions change over time. He shows that under a price cap constraint like that in expression (5), the regulated firm will often raise prices substantially on services for which demand is increasing. The firm will do so because the true impact of price increases on consumers is understated when (smaller) lagged quantities are used in place of (larger) actual quantities when calculating average price changes. In this sense, the firm is not penalised sufficiently severely for raising the prices of products for which demand is growing, and so excessive price increases are induced. See Fraser (1995) for related insights regarding the effects of exogenous cost changes on prices.

[44] Intuitively, convergence to the Ramsey optimum because constraint (6) forces the regulated firm to evaluate relative price changes each year precisely as consumers do. In particular, a price increase on a product for which demand is high reduces significantly the firm's ability to raise other prices, just as it reduces consumers' surplus significantly. (See Train (1991, pp. 156–164) for a lucid explanation of this point.)

welfare in the regulated industry to decline below the level achieved in the absence of regulation (Sappington, 1980).

5.4.3. Strategic pricing under tariff basket regulation

The regulated firm may behave strategically to relax a binding price cap constraint even when the constraint does not link allowed prices to realised costs. To illustrate the nature of the strategic behaviour, suppose constraint (5) is imposed on the firm each year. Notice that although the lagged quantity weights (q_i^{t-1}) applied to price changes in year $t(p_i^t - p_i^{t-1})$ are beyond the control of the firm in year t, the firm can influence these weights through the prices (p_i^{t-1}) it sets in year $t-1$. In particular, the firm can reduce the weight (q_i^{t-1}) applied to a price increase $(p_i^t - p_i^{t-1} > 0)$ on product i in period t by increasing the price charged for product i in period $t-1$ (which reduces q_i^{t-1}). Similarly, the firm can increase the weight applied to a price decrease in any year by reducing the price of the relevant product in the preceding year. By strategically altering the quantity weights in this manner, the regulated firm can implement gradually cumulative price changes that would be precluded by constraint (5) if the firm attempted to implement the same cumulative changes immediately. The net result of this strategic pricing behaviour can be to reduce aggregate welfare substantially (Foreman, 1995; Law, 1997).

Strategic pricing of this form can be particularly advantageous to the regulated firm when it anticipates changes in its operating environment. To illustrate, suppose the firm knows that increasing competition will soon force it to reduce substantially the price it charges for a particular product. To be sure the ultimate price decline is afforded substantial weight (q_i^{t-1}) when calculating average price changes $(\sum_{i=1}^{n} q_i^{t-1}[p_i^t - p_i^{t-1}])$, the firm can reduce the price (p_i^{t-1}) somewhat before competition compels any reduction. Doing so can provide the firm expanded freedom to raise other prices when constraint (5) is imposed (Brennan, 1989)[45].

Notice that strategic pricing of this form could be avoided if quantity (or revenue) weights were not updated annually to reflect the most recent levels of consumer demand. Instead, the output levels that prevailed at the onset of price cap regulation could be employed as weights throughout the price cap regime, for example. The drawback to using immutable weights is that, eventually, they may not reflect accurately the true impact of price changes on consumers. This drawback is deemed to be sufficiently severe in practice that weights are typically

[45] Foreman (1995) and Law (1997) analyse corresponding strategic behaviour when price changes are weighted by lagged relative revenues rather than lagged output levels. Foreman identifies conditions under which strategic weight manipulation is more problematic when relative revenue weights are employed than when quantity weights are employed.

updated annually, despite the incentives for strategic manipulation that such updating can provide.

5.4.4. *Strategic non-linear pricing*

Price cap regulation can invite an additional form of strategic pricing when non-linear pricing is permitted. To illustrate this fact most simply, suppose the regulated firm produces only one product and charges consumers for their consumption of this product according to a two-part tariff. A two-part tariff consists of a lump sum charge (or entry fee, E) and a constant usage price (p). Thus, a consumer who purchase q units of the product pays $E + pq$.

When the price charged for a product has two components (E and p), both components must be controlled in order to impose meaningful restrictions on the firm. One means of controlling both components of a two-part tariff is to restrict calculated average revenue below some specified level (p^0). Since demand is typically impossible to forecast perfectly, current demand ($q^{t-1} \equiv Q(p^{t-1}, E^{t-1})$) might be employed to approximate actual demand ($q^t \equiv Q(p^t, E^t)$) at proposed prices (p^t, E^t), as in expressions (4) and (5). Using this approximation, the restriction that calculated average revenue not exceed p^0 in year t can be written as:

$$\frac{[E^t + p^t q^{t-1}]}{q^{t-1}} \leqslant p^0. \tag{7}$$

The numerator in the fraction in expression (7) is an approximation of total revenue at prices (p^t, E^t), where lagged quantity (q^{t-1}) is employed in place of actual quantity (q^t). Dividing this expression by lagged quantity provides calculated average revenue.

Notice that expression (7) is readily rewritten as:

$$p^t + \frac{E^t}{q^{t-1}} \leqslant p^0. \tag{8}$$

Expression (8) reveals that in calculating average revenue, the proposed entry fee (E^t) is divided by lagged output (q^{t-1}). Therefore, the firm can reduce calculated average revenue by increasing lagged output, which is achieved by reducing historic usage prices (p^{t-1}). Again, then, the firm can strategically manipulate quantity weights in order to relax a binding price cap constraint. Such strategic behaviour can cause aggregate welfare to fall below the level that would be secured by requiring the firm to charge p^0 for each unit of output it sells

(Sappington and Sibley, 1992)[46]. Thus, second degree price discrimination is not always advisable under price cap regulation[47].

5.4.5. Average revenue regulation

Second and third degree price discrimination can also reduce consumers' surplus when price cap regulation restricts average revenue below a specified level and when average revenue is calculated using forecast demand (not actual lagged demand). To see why, consider the case where demand can be forecast accurately. In this case, when the firm employs linear pricing (i.e., a fixed charge per unit, p_i), the requirement that the average revenue derived from selling n products remain below a specified level (p^0) can be written as:

$$\sum_{i=1}^{n} p_i q_i \Big/ \sum_{i=1}^{n} q_i \leqslant p^0, \quad \text{where } q_i = Q(p_i). \tag{9}$$

The key feature of constraint (9), often referred to as *average revenue regulation*, is that the authorised level of average revenue applies symmetrically to all products[48]. In particular, a higher level of average revenue is not authorised for products that are more costly to produce, and lower levels of average revenue are not imposed on products that can be produced at relatively low cost. Consequently, the firm that operates under average revenue regulation will have strong financial incentive to restrict the output of products that are relatively costly to produce on the margin and increase the output of products that are less costly to produce (Bradley and Price, 1988; Waterson, 1992; Cowan, 1997b). If the products that are relatively costly to produce are also the products that are highly valued by consumers, then imposition of average revenue regulation can reduce consumers' surplus

[46] This particular form of strategic non-linear pricing might be mitigated by weighting the usage charge and the entry fee separately. For example, the entry fee might be weighted by the number of consumers, rather than their consumption levels. Cowan (1997a) analyses the drawbacks to average revenue regulation of this type when the firm produces multiple products but cannot implement non-linear pricing. He shows that strategic pricing by the regulated firm can cause welfare to fall below the level achieved in the absence of regulation.

[47] Second degree price discrimination (i.e., non-linear pricing) will not decrease either consumers' surplus or profit if it is authorised along with the mandate that consumers be afforded the option of purchasing on the uniform tariff, p^0 (Sappington and Sibley, 1992; Armstrong, Cowan, and Vickers, 1995).

[48] Since the outputs of different products are summed when calculating average revenue in constraint (9), average regulation is most appropriate in settings where the units of output of different regulated products are commensurate. Average revenue regulation has been employed in the British electric, gas, and airport industries (Armstrong et al., 1994, 1995; Cowan, 1997a).

severely, even below the level that would arise in the absence of regulation (Cowan, 1997b)[49].

In general, consumers' surplus falls when the firm is permitted to charge different prices for different products under average revenue regulation. Consumers' surplus would be higher if the firm were required to charge p^0 for each of its products (Armstrong and Vickers, 1991). Price variation harms consumers in this setting because consumers react to price variation by reducing their consumption of products with high prices and increasing their consumption of products with low prices. Average revenue declines when high prices are weighted less heavily and low prices are weighted more heavily. Therefore, the regulated firm can charge higher prices without violating constraint (9) when third degree price discrimination is permitted than when it is prohibited. These higher prices reduce consumers' surplus[50].

Reductions in the authorised level of average revenue (p^0) can also cause consumers' surplus to fall under average revenue regulation (Law, 1995). Reductions in p^0 increase the firm's incentive to reduce the sales of products that are relatively costly to produce. If consumers value these products highly, then the reduction in consumer welfare due to the reduced consumption of highly valued products can outweigh any increase in consumer welfare due to the reduction in average prices that accompanies a reduction in p^0. Thus, a more stringent price cap constraint is not always in the best interest of consumers[51].

In summary, the precise manner in which the average growth rate of prices is calculated under price cap regulation can affect the performance of the regime, at least in theory. In practice, limited knowledge of ever-changing demand and cost structures may limit the ability of the regulated firm to act strategically to relax the price cap constraint.

5.5. *The price cap constraint in practice*

Price cap regulation in the telecommunications industry is implemented most often with a variant of tariff basket regulation that differs from expression (5) in some respects, but incorporates the essential advantages and disadvantages of tariff basket regulation discussed above. This variant prohibits an average price index in year t of price cap regulation (API^t) from exceeding a specified price cap

[49] Crew and Kleindorfer (1996b) show that when total revenue, rather than average revenue, is capped, the regulated firm may be motivated to set prices above their unregulated monopoly levels. The price increases reduce sales, which in turn reduce production costs, and thereby increase the profit the firm can secure while generating the maximum allowed revenue.

[50] Armstrong, Cowan, and Vickers (1995) show that, for similar reasons, second degree price discrimination also reduces consumers' surplus when constraint (9) is imposed on the regulated firm.

[51] Kang et al. (2000) show that consumers' surplus can fall when a price cap constraint that employs exogenous weights is tightened if the firm's products are complements or substitutes. In contrast, a tighter price cap constraint always increases consumers' surplus when the firm's products are independent.

index for that year (PCI^t). Formally,

$$API^t \leqslant PCI^t \quad \text{for all } t = 1, \ldots, T, \tag{10}$$

where T is the number of years for which price cap regulation is imposed.

The actual price index in year t is an index of the prices actually charged by the regulated firm in that year, and corresponds to the term to the left of the equality in expression (1). The price cap index in year t specifies the maximum possible value for the actual price index in year t, and corresponds to the terms to the right of the equality in expression (1). The key difference between the formulations of API^t and PCI^t that follow and the terms in expression (1) is that the formulations that follow view the price indices in each year as multiples of the relevant indices in the preceding year. For reasons that will be demonstrated shortly, this inter-temporal linkage is important because it ensures that the regulated firm is not penalised forever for reducing prices below their authorised level at some point during the price cap regime[52].

Formally, the actual price index and the price cap index are defined as follows:

$$API^t = API^{t-1}\left[\sum_{i=1}^{n} r_i^r \left(\frac{p_i^t}{p_i^{t-1}}\right)\right] \tag{11}$$

$$\text{where} \quad r_i^t = \frac{p_i^{t-1} q_i^{t-1}}{R^{t-1}}, \quad \text{where } R^{t-1} = \sum_{i=1}^{n} p_i^{t-1} q_i^{t-1}; \text{ and} \tag{12}$$

$$PCI^t = PCI^{t-1}[1 + I^t - X^t], \tag{13}$$

where: n = number of products subject to price cap regulation;
p_i^t = unit price of product i in year t ;
p_i^{t-1} = unit price of product i in year $t-1$,
q_i^{t-1} = number of units of product i sold in year $t-1$;
R^{t-1} = revenues in year $t-1$ from all products subject to price cap regulation;
I^t = inflation index for year t; and
X^t = X factor for year t.

Expression (11) states that the actual price index in year t is a multiple of the actual price index in year $t-1$[53]. The relevant multiple is a weighted average of

[52] Dennis Weisman deserves the credit for this observation.

[53] The actual price index for the year prior to the start of price cap regulation is defined to be 100. The price cap index for this year is also defined to be 100.

the ratio of the prices charged by the firm in year t to the corresponding prices in year $t-1$. As expression (12) indicates, the weight (r_i^t) applied to service i in calculating this weighted average of relative prices is the fraction of total revenues generated by service i in year $t-1$.

Expression (13) states that the price cap index in each year is a multiple of the price cap index in the preceding year. The relevant multiple in year t is one plus the difference between the relevant estimate of inflation for year t (I^t) and the X factor for year t (X^t)[54].

Expressions (10), (11), and (13) imply that the critical restriction on prices in year t of price cap regulation can be written as:

$$API^t = API^{t-1}\left[\sum_{i=1}^{n} r_i^t\left(\frac{p_i^t}{p_i^{t-1}}\right)\right] \leq PCI^{t-1}[1 + I^t - X^t] = PCI^t. \tag{14}$$

Dividing the middle two terms in expression (14) by API^{t-1} provides:

$$\sum_{i=1}^{n} r_i^t \frac{p_i^t}{p_i^{t-1}} \leq \frac{PCI^{t-1}}{API^{t-1}}[1 + I^t - X^t]. \tag{15}$$

If the regulated firm sets prices at their maximum authorised levels in year $t-1$, then $API^{t-1} = PCI^{t-1}$. In this case, expression (15) can be written as:

$$\sum_{i=1}^{n} r_i^t\left(\frac{p_i^t - p_i^{t-1}}{p_i^{t-1}}\right) \leq I^t - X^t. \tag{16}$$

Expression (16) corresponds exactly to expression (1), where the rate of growth of the firm's prices ($\dot{p}$) is calculated using relative revenue weights[55].

Expression (15) reveals that if the regulated firm chooses to set prices below their maximum authorised level in some year, it is permitted to set compensating higher prices in subsequent years. Notice that if the firm sets prices below their authorised levels in year $t-1$, the price cap index will exceed the actual price index in that year (i.e., $\frac{PCI^{t-1}}{API^{t-1}} > 1$). Consequently, the firm will be authorised to

[54] Although the value of the X factor need not be the same in every year of a price cap regime, it typically is in practice.

[55] In an attempt to provide special price protection for small consumers, Oftel's price cap regulation plan for British Telecom employs relative revenue weights that reflect only the expenditures of these consumers. (Small consumers are those eighty percent of all customers who spend the least on telecommunications services.) When relative revenue weights do not reflect the expenditures of large customers, price reductions provided exclusively to large customers do not automatically authorise price increases on services sold primarily to small customers.

implement price increases in year t that exceed the relevant difference between the inflation index and the X factor ($I^t - X^t$). Because it is not penalised forever for setting particularly low prices one year, the firm will be more willing to reduce prices below authorised levels some years, which can be advantageous for consumers.

5.6. Additional restrictions on individual price levels

Although price cap regulation limits the rate at which regulated prices can increase on average, it does not necessarily preclude substantial changes in individual prices. If the aggregate price constraint (as captured in expression (1), for example) is the only constraint that is imposed, then the price of any one service can rise dramatically as long as the prices of other services decline by a sufficiently large amount. Intense opposition to a price cap plan can emerge if it permits the prices of certain key services to rise too rapidly. Therefore, price cap plans often incorporate additional restrictions on the rate at which the prices of selected services can change. These additional restrictions typically take the form of price ceilings, price floors, and pricing bands.

5.6.1. Individual price ceilings

Ceilings are often imposed on the prices of certain regulated services. In the telecommunications industry, for example, the price of basic local telephone service for residential customers is often frozen for some period of time, and then permitted to rise only slowly thereafter. Restrictions of this type commonly reflect political realities, even if they are at odds with economic principles. Regulators are often compelled to please their constituents, particularly their most vocal constituents. If highly visible rates that most customers must pay (such as basic local telephone service rates) rise too quickly, widespread consumer dissatisfaction can arise. To prevent or mitigate such dissatisfaction, separate restrictions on key service rates are common, even if the restrictions preclude prices from rising toward relevant production costs.

5.6.2. Price floors and imputation

Price floors on individual service rates are also common under price cap regulation. Typically, the price of a regulated service is not permitted to fall below the regulated firm's long run incremental cost of supplying the service. This restriction prevents the firm from establishing non-compensatory rates in order to deter the operation of more efficient suppliers[56].

Price floors in excess of long run incremental cost are often imposed when a regulated firm supplies an essential input to alternative producers and also

[56] See, for example, Vickers (1997).

supplies a retail product in direct competition with these producers. To illustrate why, suppose a regulated telecommunications firm supplies basic network access to competing unregulated providers of long distance telephone service. Each unit of access costs the regulated firm c^a to supply, and the firm charges p^a for each unit of access it supplies. Suppose the regulated firm also sells long distance telephone service to customers. Each unit of long distance service requires one unit of access, and costs the firm $c^a + c^d$ in total to supply. If the regulated firm were free to reduce the price it charged for long distance service to the level of the incremental cost of providing this service $(c^a + c^d)$, the firm might be able to disadvantage its competitors by implementing a price squeeze. A price squeeze would occur if the regulated firm set the price it charges competitors for network access (p^a) above $c^a + c^d$, and then reduced its price for long distance service toward $c^a + c^d$. Such pricing would ensure that competitors could not profitably match the price that the regulated firm established for long distance service.

To preclude such price squeezes, when price floors are established, the costs that the regulated firm imposes on its retail competitors are commonly treated as costs that the regulated firm incurs itself. To illustrate, the regulated price floor for long distance service in this setting would be the sum of the price charged to competitors for access (p^a) and the incremental cost of supplying long distance service, given access (c^d). By treating the costs imposed on competitors as costs incurred by the regulated firm in this manner (a procedure knows as *imputation*), prize squeezes can be avoided, thereby securing a more level playing field for retail competition[57].

5.6.3. *Pricing bands*

Sometimes price floors are established at levels that exceed relevant incremental cost (including imputed cost) and are coupled with price ceilings. The result is a pricing band that restricts the extent to which the price for a service can be either raised or lowered. Typically, pricing bands specify a maximum percentage by which the price for a service can be raised or lowered annually from initial levels. For example, the pricing band may prohibit price increases or decreases of more than 10 percent in any year.

The primary purpose of pricing bands is to slow the rate at which prices change under price cap regulation. In this sense, pricing bands can add stability to a price cap plan. They do so, however, at the cost of limiting the ability of incumbent suppliers to move prices toward relevant production costs and to respond to competitive pressures.

[57] Notice that price cap regulation does not necessarily preclude price squeezes. Aggregate restrictions on overall price levels admit substantial increases in some prices (e.g., for network access) provided they are offset by decreases in other prices (e.g., for long distance service). Therefore, imputation rules and/or distinct baskets of wholesale and retail services (see Section 5.7) may be advisable if global price cap regulation, as proposed by Laffont and Tirole (1996), is adopted.

5.7. The number of baskets

Price changes on individual services can also be limited by placing services into distinct baskets, each with its own restriction on average price levels. The greater the number of distinct baskets into which services are placed, the less pricing flexibility the regulated firm is afforded.

To illustrate, telecommunications services purchased primarily by residential customers might be placed in one basket, while services purchased primarily by business customers might be placed in a different basket. By separating services in this manner, the regulated firm is not afforded the opportunity to offset substantial price increases on services sold to residential customers with price decreases on services sold to business customers, for example. Thus, the use of separate baskets is another means of providing price protection for particular groups of customers or for consumers of particular groups of services. Such protection can be particularly important when the regulated firm faces different competitive pressures on different groups of products. Separate baskets can reduce the incentive the regulated firm might otherwise have to set the prices of services facing intense competition below incremental production cost and compensate for the ensuing reduction in earnings by increasing the prices of services for which competition is less intense (Armstrong and Vickers, 1993)[58].

Separate baskets for wholesale and retail services can also limit incentives for price squeezes. When reductions in the prices of retail services do not authorise increases in the prices charged to competitors for essential inputs, price squeezes may become less profitable for, and thus less attractive to, a regulated supplier of both wholesale and retail services.

One disadvantage of placing services in distinct baskets and imposing a separate price cap constraint on each basket is that the separation limits the regulated firm's ability to realign prices to more closely approximate production costs. Consequently, establishing distinct baskets and separate caps can prevent the regulated firm from setting prices that converge to Ramsey levels (as it might otherwise do when a constraint like expression (6) is imposed)[59].

[58] It is generally preferable to replace regulatory control with the discipline of competition when competition provides adequate protection for consumers. In practice, though, it is often difficult to determine precisely when adequate, sustainable competitive pressures have developed. Furthermore, the regulated firm may lobby to include competitive services in the basket of price-capped services, since their inclusion can authorise the firm to raise prices on non-competitive services to offset price reductions on competitive services. Thus, for a variety of reasons, regulatory control can persist beyond the point at which it is no longer warranted.

[59] This is, in part, the reason that Laffont and Tirole (1996) recommend global price cap regulation, under which all of the regulated firm's services (both retail and wholesale services) are placed in a single basket and controlled with a single aggregate constraint.

5.8. Service quality

Although price cap regulation plans typically focus on regulating service prices, they also commonly regulate service quality. Service quality regulation is often thought to be particularly important under price cap regulation because the enhanced incentives for cost reduction that price caps provide can render certain reductions in service quality particularly profitable. On the other hand, by enabling regulated suppliers to retain more of the revenue that they generate in the market place, price cap regulation may render increases in certain dimensions of service quality particularly profitable. Thus, it is not clear *a priori* whether price cap regulation provides enhanced or diminished incentives for service quality relative to rate-of-return regulation.

In settings where a standard price cap regulation plan and competition together are thought to provide inadequate incentives for service quality, suitable modifications of price cap plans can be implemented. For example, key dimensions of service quality (e.g., the time a customer must wait to receive assistance with a billing problem) can be identified and a target level of performance can be specified for each dimension. Suitable financial rewards and penalties can then be implemented for service quality that exceeds or falls short of specified targets[60]. Rewards and penalties can be delivered in a variety of ways. For instance, financial rewards can be effected by raising the cap on allowed prices if service quality surpasses specific thresholds. Financial penalties might be imposed either by reducing the authorised overall rate of price increases or by requiring the regulated firm to compensate customers directly for the inconvenience they incur when the quality of service they receive falls below specified levels. For instance, local telephone companies may be required to pay customers a fixed daily fee and/or provide them with temporary wireless telecommunication service if wireline service is not installed in a timely fashion[61].

5.9. Conclusions

Although the over-arching constraint on prices under price cap regulation (constraint (1)) is relatively simple conceptually, its implementation raises a host of complicating considerations, as do the other constraints (on individual service prices, quality, etc.) that price cap plans typically impose. In practice, pure price cap plans are often modified in a variety of ways in an attempt to improve their performance. Some common modifications are discussed in the next section.

[60] Berg and Lynch (1992) and Lynch et al. (1994) explain the merits of basing rewards and penalties on an index of service quality performance measures rather than on individual service-specific performance measures.

[61] See Rovizzi and Thompson (1992).

6. Hybrid regulatory plans

Pure price cap regulation severs the link between allowed prices and realised costs. Therefore, as noted above, it can result in extremely high or extremely low earnings for the regulated firm. Regulators are typically averse to unusually high earnings, in part because constituents may view high earnings as a sign that the regulator favoured the firm unduly when designing the price cap regime. Particularly low earnings can also be troubling for both the regulator and the firm, in part because low earnings can threaten the firm's ability to attract capital and provide high quality service to customers on a continuing basis.

There are a variety of ways in which price cap regulation plans can be structured to avoid extreme profit levels for prolonged periods of time. As noted in Section 5, a short time span between reviews of the price cap plan can be instituted, and the reviews can be structured primarily to align prices and costs. Alternatively, or in addition, mid-term modifications of the cap (Z factors) can be permitted for major unanticipated events that are beyond the control of the regulated firm.

This section analyses in more detail three of the additional modifications of pure price cap regulation described in Section 2 that can help to limit extreme profit variation during a price cap regime. The three modifications are earnings sharing, revenue sharing, and regulatory options, all of which have been employed to varying degrees in the telecommunications industry.

6.1. Earnings sharing plans

Recall from Section 2 that earnings (or profit) sharing plans require the regulated firm to share with its customers a portion of the earnings that it generates in the market. A typical earnings sharing plan is illustrated in Figure 2. Under that plan, the firm must deliver to its customers half of each additional dollar of earnings it generates when earnings constitute a rate of return between 14 percent and 16 percent. Higher incremental earnings accrue entirely to customers. As it does under pure price cap regulation, the firm retains all incremental earnings when its rate of return is between 10 percent and 14 percent.

Incremental earnings can be shared in a variety of ways. Two common procedures are direct payments to customers (typically in the form of a credit on their bill) and changes in the prices of key services (such as basic local telephone service). Direct payments to customers have the advantage of demonstrating clearly to consumers the benefits that the incentive regulation is delivering to them (Sappington and Weisman, 1994b; 1996a, p.186). Price adjustments can offer the advantage of aligning prices more closely with costs, thereby increasing total surplus in the industry[62].

[62] Sometimes, though, shared earnings are employed to move the prices of popular services – like basic local telephone service – further below their incremental cost of production.

6.1.1. The fundamental trade-off

Regardless of the manner in which earnings are shared, the requirement to share earnings with customers alters the regulated firm's incentives to minimise operating costs and increase revenues. When the firm bears substantial unmeasured costs in improving operating efficiency, the firm's incentives to improve its operations can be dulled by earnings sharing[63]. This is the fundamental trade-off associated with earnings sharing. Earnings sharing reduces the incidence of extreme earnings but dulls the firm's incentives for outstanding performance[64].

If earnings are shared by moving prices closer to realised cost, then a small amount of earnings sharing increases aggregate welfare (the sum of consumers' surplus and profit). This is because the losses from diminished incentives to minimise production costs are small relative to the gains from better aligning prices and costs when small amounts of earnings sharing are introduced (Lyon, 1996). More pronounced earnings sharing can reduce welfare by reducing substantially the firm's incentive to reduce its operating costs. However, even though pronounced earnings sharing can reduce aggregate welfare, it can increase consumers' surplus by transferring realised surplus from the firm to its customers. Therefore, as noted in Section 4, more pronounced earnings sharing is generally optimal the more highly consumers' surplus is valued relative to profit. Earnings sharing can be particularly valuable in securing surplus for consumers when the regulator faces considerable uncertainty about the firm's ability to reduce operating costs and when costs are initially high, so the potential for cost reduction is substantial (Schmalensee, 1989; Lyon, 1996)[65]. In contrast, when the firm's ability to reduce costs is thought to be particularly pronounced, less profit sharing is generally advisable in order to motivate the firm to deliver a substantial amount of cost-reducing effort (Lyon, 1996).

6.1.2. The nature of sharing arrangements

Although most earnings sharing plans that are implemented in practice resemble the plan depicted in Figure 2, basic economic principles suggest that a different plan would provide stronger incentives for the firm to pursue substantial reductions in operating costs. A key feature of the plan in Figure 2 is the fact that when the firm's earnings exceed the target rate of return (12 percent), the fraction of incremental earnings awarded to the firm declines with the level of earnings it generates. In particular, the firm receives first all, then half, and finally none of its

[63] Relevant unmeasured costs include those associated with increased managerial effort or a more stressful working environment, for example.

[64] If the earnings moderations secured by an earnings sharing provision convinces the regulator to implement price cap regulation for a longer period of time, then earnings sharing may result in a less pronounced diminution of the firm's incentives for outstanding performance.

[65] Gasmi et al. (1994) report that simple profit sharing plans coupled with downward pricing flexibility for the firm can often closely approximate optimal incentive structures.

incremental earnings as market returns rise toward 14 percent, then toward 16 percent, and finally above 16 percent. When higher returns stem primarily from realised cost reductions, this pattern of earnings sharing promises smaller incremental rewards for additional cost reduction the greater the amount of cost reduction already achieved. If it becomes increasingly more difficult to achieve further cost reduction the greater the level of cost reduction already achieved, then the diminishing incremental rewards promised by an earnings sharing plan like the one depicted in Figure 2 will fail to provide strong incentives for particularly extensive cost reductions. An earnings sharing scheme that promised a greater share of incremental earnings to the firm the higher the level of earnings it achieves would better serve this purpose (Blackmon, 1994, 60–62; Lyon, 1996). However such schemes run the risk of delivering profits to the firm that are deemed to be unduly large, and so, in part for this reason, are uncommon in practice[66].

6.1.3. Investment incentives

In addition to reducing the regulated firm's incentive to reduce operating costs and increase revenues, earnings sharing plans can provide undesirable investment incentives to the firm. To see why, consider again the earnings sharing plan illustrated in Figure 2. When the firm generates earnings in the market that constitute a return of approximately 10 percent, the firm faces asymmetric returns from additional investment projects. If a project fails and reduces returns below 10 percent, the firm bears only half of the associated costs. In contrast, the firm reaps all of the gains if the project increases returns above 10 percent. When it enjoys the full upside potential of projects but bears only a fraction of their downside risk, a firm may find it profitable to undertake projects with low (even negative) expected net present discounted value (NPDV). A firm may also find it profitable to undertake projects that are unduly risky (Blackmon, 1994, 66–68).

For analogous reasons, the regulated firm operating under the earnings sharing plan in Figure 2 may choose not to pursue projects with positive expected NPDV when its current earnings constitute a return of approximately 14 percent. At this point, the firm can bear the full cost of a failed project but receive at most half of the returns from a successful project. Thus, when they incorporate abrupt changes in the marginal returns to improved performance, earnings sharing plans can induce undesired investment behaviour[67]. One way to limit the extent of this behaviour is to include numerous small changes, rather than a few large changes,

[66] Davis (2000) reports that under the earnings sharing plan implemented for San Diego Gas and Electric in 1999, the company is afforded a larger fraction of incremental earnings as earnings increase between 25 and 300 basis points above the approved rate of return.

[67] Earnings sharing plans can improve investment incentives when regulators are otherwise inclined to force the firm to bear the downside risk of failed investments and limit the firm's returns from successful investments (Lyon, 1995).

in the rate at which earnings are shared between the regulated firm and its customers (Blackmon, 1994, 67–68).

6.1.4. Cost shifting

Earnings sharing schemes can also provide incentives for the regulated firm to shift costs from one year to the next. To see why, suppose a regulated firm is operating under the earnings sharing plan depicted in Figure 2. Also suppose the firm is currently generating a 9 percent return in the market, but anticipates earning a 12 percent return next year. In this situation, the firm can gain financially by shifting costs from next year to the current year, perhaps by accelerating routine maintenance expenditures, for example. Since the firm bears only half of any realised cost increase when its earnings provide a return of 9 percent but benefits by the full amount of any realised cost reduction when its earnings provide a 12 percent return, the firm gains financially from shifting costs intertemporally in this manner.

6.1.5. The regulator's incentives

Earnings sharing plans can also affect the incentives of the regulator, just as they affect the incentives of the regulated firm. In particular, earnings sharing plans can alter the regulator's incentive to disallow realised production costs when calculating earnings and to promote entry into the regulated industry.

Consider, first, the regulator's incentive to disallow costs. If a cost incurred by the regulated firm is not counted as such when calculating the firm's earnings, then measured earnings increase. Consequently, if the firm's earnings are in the range where nontrivial earnings sharing is mandated (e.g., if they constitute a rate of return between 14 percent and 16 percent under the plan depicted in Figure 2), the regulator can secure a greater direct payoff for consumers by declaring some costs to have been incurred imprudently, and thereby disallow those costs[68]. The regulator who values consumers' surplus more highly than profit will find it attractive to disallow costs under earnings sharing plans unless doing so raises the firm's cost of capital by more than the corresponding direct gains secured for consumers. By increasing the regulator's incentive to disallow costs in this manner, earnings sharing plans can induce the regulated firm to undertake too little surplus-enhancing investment (Sappington and Weisman, 1994a; 1996c).

Now consider the impact of earnings sharing plans on the incentives of regulators to promote entry into the regulated industry. Recall that under pure price cap plans, prices do not vary with the earnings of the regulated firm. In particular, the regulator is under no obligation to raise prices in the regulated

[68] See Howe (1983; 1985), Kolbe and Tye (1991), Lyon (1991; 1992), and Encinosa and Sappington (1995), for example, for analyses of regulator prudence policy.

industry as the firm's earnings fall. This fact may encourage the regulator to facilitate entry into the industry in order to secure even lower prices for consumers. The regulator may be more reluctant to encourage entry when an earnings sharing plan is in place because the plan can obligate the regulator to raise industry prices in order to mitigate any major impact of entry on the earnings of the incumbent firm[69]. Since earnings sharing plans can better align the interests of the regulator and the regulated firm in this manner, the firm may ensure higher *ex post* returns by agreeing *ex ante* to share earnings with customers (Weisman, 1994).

6.1.6. Summary

In summary, earnings sharing offers a host of benefits and costs[70]. It can help to avoid extreme profit realisations, but only at the cost of dulling the firm's incentives for cost reduction and distorting its investment incentives. Earnings sharing plans can also affect the regulator's incentives to disallow costs and promote entry into the regulated industry. As noted in Section 3, the recent trend in the U.S. telecommunications industry is away from earnings sharing plans and toward pure price cap plans. This trend may reflect in part reduced uncertainty about the ability of incumbent telecommunications suppliers to increase their earnings when afforded substantial incentive to do so.

6.2. Revenue sharing plans

Profit variation can also be dampened under price cap regulation plans by imposing revenue sharing rather than earnings sharing. As illustrated in Figure 3, a typical revenue sharing plan requires the regulated firm to share with its customers incremental revenue in excess of a specific threshold level.

Because they mandate the sharing of revenues, not profits, revenue sharing plans do not reduce the regulated firm's incentive to minimise its operating costs. In addition, unlike earnings sharing plans, revenue sharing plans provide no incentives for the regulator to disallow production costs incurred by the firm. It is also not necessary to measure earnings under price cap regulation plans that incorporate revenue sharing. Therefore, revenue sharing plans avoid some of the key drawbacks that plague earnings sharing plans.

[69] Lehman and Weisman (2000) provide empirical evidence of this effect.

[70] An additional cost of an earnings sharing plan is the resources it consumes in measuring earnings. Such measurement is generally not straightforward. Earnings can be difficult to measure in part because operating costs generally include depreciation costs, and depreciation costs can very with the estimated value of the firm's assets, which in turn vary with regulatory policy (Mayer and Vickers, 1996). Earnings sharing plans can also provide the regulated firm with inappropriate incentives to diversify into unregulated markets and to shift costs across markets, depending upon how cost allocation rules are structured (Weisman, 1993; Chang and Warren, 1997).

Revenue sharing plans are not without their own potential drawbacks, however[71]. Most importantly, because the firm bears the full cost of generating revenues above the specified threshold but receives only a fraction of the associated benefit, the firm will generally have insufficient incentive to generate these incremental revenues. Such insufficient incentive can manifest itself in a variety of ways. For instance, the regulated firm may not install the capacity required to satisfy all demand that is expected to materialise. In addition, the firm might reduce the quality of service it provides when reduced service quality reduces operating costs. The firm might also alter relative prices. It might, for example, lower prices on services with inelastic demand and raise prices on services with elastic demand. Doing so can increase the firm's earnings by reducing aggregate output and thus operating costs (Sappington and Weisman, 1996c).

The merits of revenue sharing vary according to the environment in which it is imposed. In an environment where the potential for cost reduction is thought to be substantial and where realised demand is largely beyond the control of the regulated firm, revenue sharing may be preferable to earnings sharing as a means of avoiding extreme profit realisations. In contrast, where service quality is highly valued, costly to provide, difficult to monitor, and serves as a primary determinant of consumer demand, revenue sharing may not be the best way to limit variation in realised profit.

6.3. Regulatory options

The choice between different regulatory plans such as rate of return regulation, pure price cap regulation, earnings sharing regulation, revenue sharing regulation, and others, can be a difficult one. The ideal regulatory plan depends upon the environment in which it is implemented, and perfect information about the regulated industry is almost never available. Fortunately, a regulator need not always select the single most appropriate regulatory plan. Instead, the regulator can allow the regulated firm to choose one regulatory regime from a carefully designed menu of regimes. Doing so can often secure the key advantages of the regulatory plans without incurring all of their disadvantages. It can also employ the regulated firm's superior knowledge of its operating environment to select a plan that is advantageous for consumers.

To illustrate this point most simply, consider the following example that is drawn from Sappington and Weisman (1996a, 158–162). The example illustrates how, by affording the firm a choice between rate of return regulation (RORR) and pure price cap regulation (PCR), the regulator can secure a higher level of expected consumers' surplus than if the regulator limited himself to implementing

[71] As noted above, Crew and Kleindorfer (1996b) show that when a binding ceiling is placed on the firm's total revenue, the firm may set prices above their monopoly levels in order to secure the maximum authorised revenue while limiting output, and thus cost.

either RORR or PCR[72]. The gain in consumers' surplus arises because when RORR is made available as an option, the regulator can offer the firm a more demanding price cap plan without risking financial distress for the firm.

The example has the following features. The regulated firm knows precisely the maximum possible improvement in its productivity growth rate under price cap regulation. Call this maximum level of improvement x_m. The regulator does not share the firm's knowledge of x_m. Thus, if it were to impose PCR, the regulator would be forced to choose an X factor, denoted x, not knowing with certainty whether the selected X factor, x, is smaller than x_m, and thus readily achieved by the firm, or whether x exceeds x_m, so that the price cap plan will force the firm to suffer financial insolvency[73]. The regulator's beliefs about the firm's capabilities, x_m, are represented by the distribution function $F(\tilde{x})$, which denotes the probability that x_m is less than or equal to any specified X factor, $\tilde{x}$.

RORR is initially practised in the industry, and consumers receive net benefits $\bar{B}$ under RORR. If the regulated firm becomes insolvent, consumer benefits fall to the lower value, $\underline{B}$. If the firm is solvent under PCR, consumer benefits rise to $[1 + x]\bar{B}$. Thus, consumer benefits are higher the more ambitious the productivity offset, x, as long as the regulated firm is solvent.

To calculate the gains that arise when PCR is offered as an option rather than mandated, it is useful to compare two distinct problems. In the first problem, the regulator chooses the value for the X factor to maximise expected consumer benefits when the regulated firm is required to operate under PCR. The second problem adds RORR as an option. Formally, these two problems are the following:

Problem 1. Mandatory Price Cap Regulation.

$$\underset{x}{\textit{Maximize}} \quad F(x)\underline{B} + [1 - F(x)]\bar{B}[1 + x].$$

Problem 2. Optional Price Cap Regulation.

$$\underset{x}{\textit{Maximize}} \quad F(x)\bar{B} + [1 - F(x)]\bar{B}[1 + x].$$

As their formulations reveal, these two problems differ only in the level of consumer benefits that are generated if the specified X factor, x, exceeds the

[72] More generally, a regulator might design an entire menu fo regimes without restricting *a priori* the nature of permissible regimes, and allow the firm to choose one regime from the menu. In such a setting, the menu of regimes will typically include a variety of different earnings sharing plans and one plan with no earnings sharing (e.g., a price cap regulation plan). See, for example, Baron and Myerson (1982), Laffont and Tirole (1986; 1993), and Lewis and Sappington (1989).

[73] In practice, insolvency is unlikely to result immediately. Instead, the firm that cannot achieve productivity gains of at least x may be forced to reduce service quality and allow its capital stock to deteriorate in the short run. As time progresses, the firm will have difficulty attracting new capital, and may eventually be forced to terminate operations. This entire process will be referred to as 'insolvency' in the ensuing discussion.

maximum potential productivity improvement of the firm. When PCR is mandated (Problem 1), the firm will be insolvent when $x > x_m$, and the smaller level of benefits, $\underline{B}$, will result. When RORR is retained as an option (Problem 2), the higher benefits, $\bar{B}$, are generated when $x > x_m$. Recall that $F(x)$ is the regulator's assessment of the likelihood that x_m is less than x, so the firm is expected to opt for RORR with probability $F(x)$ in Problem 2. With the complementary probability, $1 - F(x)$, the firm is expected to choose PCR, resulting in consumer benefits of $\bar{B}[1 + x]$.

It is straightforward to verify that the regulator will select a more demanding productivity offset (x) when PCR is offered as an option than when its adoption is mandated[74]. Consumers suffer less when RORR rather than insolvency arises if the firm cannot achieve the productivity gains required under PCR. Consequently, the regulator is more willing to implement a productivity target that exceeds the firm's capabilities. It is also readily shown that a higher level of expected consumer benefits results when the higher productivity offset is selected in Problem 2.

The magnitude of the expected gain from introducing price cap regulation as an option rather than mandating its adoption depends on the nature of the uncertainty about the firm's ability and on the loss incurred if the regulated firm becomes insolvent. To illustrate, suppose the regulator is certain that the firm's maximum potential productivity gain (x_m) is between 1 and 5 (percent). Within this range, however, all possible realisations are equally likely[75]. It can be shown that in this setting, the regulator will set $x = 2.5$ when PCR is optional. He will set $x = 2 + .5\underline{B}/\bar{B} < 2.5$ when PCR is mandated. The corresponding levels of expected consumer benefits are $2.56\bar{B}$ when PCR is optional, and $.25[\underline{B} + 9\bar{B} + .25(\underline{B})^2/\bar{B}]$ when it is mandatory. Therefore, if, for example, $\underline{B} = .25\bar{B}$, so the consumer benefits generated when the firm is insolvent are one quarter of the corresponding benefits that are secured under RORR, the regulator will set $x = 2.5$ when PCR is optional and $x = 2.125$ when it is mandated. Expected consumer benefits will be approximately eleven percent higher when PCR is offered as an option rather than mandated. Thus, the gains from providing choices to the firm when implementing incentive regulation plans can be substantial.

The gains from options can be even larger when regulation is being designed for more than a single regulated firm. The ideal regulatory plan for one firm may be largely inappropriate for another firm. Consequently, by affording multiple regulated firms a choice among regulatory options, a regulator can treat all firms identically *ex ante* while securing *ex post* differences in regulatory policies that best serve consumers. As noted in Section 2, the Federal Communications

[74] Formally, if x_1 denotes the (unique, interior) solution to Problem 1, x_1 is given by $\bar{B}[1 - F(x_1)] = [\bar{B}[1 + x_1] - \underline{B}]F'(x_1)$. Also, if x_2 denotes the (unique interior) solution to Problem 2, then x_2 is given by $\bar{B}[1 - F(x_2)] = \bar{B}x_2F'(x_2)$.

[75] Formally, $F(x) = .25[x - 1]$ *for all* $x \in [1, 5]$.

Commission afforded different options to the major local exchange carriers in the United States. By doing so, it was able to better tailor regulatory policy to the different operating circumstances of the different carriers.

Although a regulator can secure gains for consumers by allowing a regulated firm to choose among regulatory options, ongoing unfettered choice among options is not always advisable. Such choices can invite strategic behaviour on the part of the firm. To illustrate, suppose the firm were permitted to choose freely between a PCR plan and a RORR plan each year. Then the firm might find it profitable to alternate between the two plans each year, and shift expenditures from years in which it operates under PCR to years in which it operates under RORR. By doing so, the firm can increase profit under PCR without reducing profit under a RORR regime that matches allowed revenues to measured costs. To limit the firm's ability to impose excessive costs on consumers in this manner, the firm might only be permitted to choose RORR after selecting PCR if profits are sufficiently low for a sufficiently long period of time under PCR.

6.4. Conclusions

This chapter has reviewed some of the ways in which pure price cap regulation plans can be modified in order to retain the central advantages of price cap regulation while avoiding some of its main disadvantages. None of these modifications is without its own potential drawbacks, though, and some can serve to blur the practical distinctions between rate of return regulation and its alternative.

Having reviewed the primary alternatives to rate of return regulation that are employed in practice, the discussion turns now to the measured impact of incentive regulation plans.

7. The impact of incentive regulation

The purpose of this section is to summarise existing knowledge regarding the impact of incentive regulation on performance in the telecommunications industry[76]. Performance has many dimensions, including prices, operating costs, network modernisation, productivity, profit, service quality, and telephone subscription rates. Each of these dimensions is examined in turn. Most of the empirical studies to date examine the effects of incentive regulation in state telecommunications markets in the United States. This is because the significant variation in regulatory policy across the fifty states constitutes a natural experiment that lends itself to econometric analysis. There is a great need to conduct corresponding empirical analyses in other countries around the world.

[76] The discussion in this section draws heavily from Kridel et al. (1996). See Abel (2000) for a complementary review of the empirical literature on the effects of price cap regulation.

7.1. Prices

The evidence regarding the impact of incentive regulation on prices is mixed. Mathios and Rogers (1989) find that AT&T set prices for intrastate, interLATA telephone calls that were approximately 7 percent lower in states where it was afforded limited pricing flexibility than its states where it faced strict rate of return regulation, holding other factors constant[77]. Although this early finding may suggest that incentive regulation encourages price reductions, Mathios and Rogers' finding may also be testimony to the effects of competition. If state regulators granted AT&T expanded pricing flexibility precisely in those states where it faced the greatest pressure from competitors, then the observed price reductions may be due more to competition than to incentive regulation[78]. The authors did not control for the possible endogeneity of the regulatory regime[79].

Kaestner and Kahn (1990) provide some corroborating evidence. They report that between 1986 and 1988, AT&T's intrastate toll prices were lower by 18 percent in states where the company was afforded pricing flexibility than in states where it faced strict rate of return regulation. Even more pronounced price reductions were observed in states where AT&T's toll prices were largely deregulated. Kaestner and Kahn employ proprietary data on non-AT&T market share in an attempt to separate the effects of competition from the effects of incentive regulation. This procedure is admirable, but perhaps not entirely successful. Increased competition may have led to relaxed regulation in the time period under study, so the observed reduction in prices may not be due entirely to incentive regulation. The authors also explain that the observed decline in deregulated prices may be due to 'demonstration effects,' i.e., in an attempt to demonstrate the merits of deregulation to all regulators, AT&T intentionally set prices below profit-maximising levels[80].

In a more recent study that examines a longer time period (1980–1991), Tardiff and Taylor (1993) find that the intraLATA toll prices set by the regional Bell Operating Companies (RBOCs) in the United States are approximately 5 percent lower on average under incentive regulation than under rate

[77] Recall that LATAs are geographic areas in the U.S. called local access and transport areas, and that inter-exchange carriers can provide telecommunications services that cross LATA boundaries (i.e., interLATA services), but the major local exchange carriers (the regional Bell Operating Companies or RBOCs) presently cannot.

[78] Abel (1999) and Banker et al. (1995, 1996) find that growing competition has compelled major price reductions in the U.S. telecommunications industry in recent years. Knittel (1997) also provides evidence which suggests that competition may have led to lower local telephone rates between 1984 and 1993.

[79] As noted above, Donald and Sappington (1995; 1997) examine the determinants of the choice of regulatory regime.

[80] See Sappington and Weisman (1996b) for a more complete discussion of demonstration effects and other factors that merit consideration when attempting to assess the impact of incentive regulation on industry performance.

of return regulation. The largest decline in intraLATA toll prices was observed under earnings sharing regulation, in part, perhaps, because price reductions are the means by which relevant earnings are shared with customers. Prices were actually higher under deregulation, price cap regulation, and rate case moratoria than under rate of return regulation. Higher toll prices under these regimes may have served to offset lower basic service rates. Basic local service rates are often the most visible and the most politically sensitive rates. Consequently, regulated firms may either have been required to keep these rates low as a pre-condition for incentive regulation, or they may have kept the rates low voluntarily in order to garner increased support for continued incentive regulation.

Blank et al. (1998) report that intraLATA toll prices charged by the RBOCs were approximately 12 percent higher under incentive regulation than under rate of return regulation at the end of 1991, after controlling for other possible causes of price variation. In particular, the authors employ several proxies for competition, including: (1) whether intraLATA toll competition is permitted in the state; and (2) the fraction of total access lines in the state that are business access lines (since competition for business customers is often particularly intense). The authors find that competition tends to reduce intraLATA toll rates by almost as much as incentive regulation increases them. The authors suggest that the expanded pricing flexibility that often accompanies incentive regulation may enable the regulated firm to raise intraLATA toll prices in return for lowering the prices of other services.

The price for basic local telephone service appears to have declined modestly under some forms of incentive regulation. Although Tardiff and Taylor (1993) report no significant decline in these prices under incentive regulation, broadly defined, prior to 1992, Crandall and Waverman (1995) find these prices to be lower by approximately 10 percent under price cap regulation than under rate of return regulation. The authors examine basic local service rates between 1987 and 1993. Ai and Sappington (1998) derive similar conclusions in their analysis of U.S. local service rates between 1991 and 1996.

Magura (1998) reports the largest estimated impact of incentive regulation on basic local service rates. Using data on U.S. basic local service rates between 1987 and 1994, Magura finds that incentive regulation is associated with a 17 percent decline in rates. The author attributes this decline to a reduction in operating costs under incentive regulation.

Undoubtedly, basic local service rates are determined by political and strategic factors as well as economic factors. Notice, for example, that basic local service rates in Nebraska have remained largely unchanged for twelve years, despite being largely deregulated in 1987. It seems likely that U.S. West intentionally kept basic local service rates far below profit maximising (and authorised) levels in Nebraska in order to foster support for reduced regulatory control in other jurisdictions where it operates.

Braeutigam et al. (1997) provide an important extension of this line of research. The authors point out that not all incentive plans within a specified category are identical. For instance, some price cap regulation plans require the regulated firm to reduce basic local service rates at the outset of the price cap regime, while others do not. Similarly, some plans require particular types of infrastructure investment and network modernisation, while others do not. Braeutigam et al. Find little impact of incentive regulation, broadly defined, on basic local service prices in the United States between 1987 and 1993. However, prices are significantly lower under plans that require initial price reductions, and are significantly higher under plans that impose substantial investment requirements. This work illustrates the importance of defining 'incentive regulation' precisely when examining its impact empirically.

Armstrong et al. (1994) report fairly dramatic reductions in British Telecom's toll prices under price cap regulation between 1984 and 1993. For example, inflation-adjusted peak-period toll calls fell by more than 60 percent during the period. In contrast, the basic monthly access charge for local service (the line rental rate) increased in real terms by approximately 5 percent. The authors point out that competitive pressures likely played a significant role in the observed decline in toll prices. However, they do not attempt to isolate the impact of competition, price cap regulation, and other factors econometrically[81].

7.2. Operating costs

Proponents of incentive regulation stress its ability to foster cost reduction by the regulated firm. When prices are permitted to diverge from realised production costs, the regulated firm can benefit financially when its production costs fall. Consequently, incentive regulation may induce the regulated firm to undertake the actions required to reduce operating costs.

To date, however, the relevant empirical evidence on this issue is mixed. Shin and Ying (1993) find that incentive regulation is associated with a slight (1 percent) increase in operating costs in the U.S. telecommunications industry between 1988 and 1991. Ai and Sappington (1998) find no link between incentive regulation and operating costs between 1991 and 1996. While Ai and Sappington simply regress total cost on historical cost and other possible determinants of cost, Shin and Ying estimate the parameters of a multi-product translog cost function. They do so using observations for 50 U.S. local exchange companies (including the RBOCs) between 1988 and 1991. The authors disaggregate total operating costs into capital costs and other costs.

[81] Wolak (1996b) examines average revenue per access line for incumbent operators in selected U.S. states. He finds significantly lower average revenue in states that have typically promoted competition in their telecommunications industries. Wolak does not control for differences in incentive regulation plans in his five-state sample.

As noted above, Magura (1998) finds that U.S. basic local service rates are approximately 17 percent lower under incentive regulation than under rate of return regulation. Under the assumption that basic local service rates are priced residually (i.e., they are set as low as possible while generating a fair return for the firm, after setting prices for other services to maximise net revenues), Magura suggests that the lower local service prices may reflect pronounced declines in fixed operating costs under incentive regulation.

The diverse findings in the literature, including the mixed results regarding the impact of incentive regulation on operating costs, may be explained by a variety of factors. First, lasting effects of incentive regulation on costs may take more time to develop than is reflected in available time series data. Second, firms may produce a different array of products and services under incentive regulation, which can affect production costs. Third, to the extent that incentive regulation is implemented when competition is more pronounced and to the extent that competition raises operating costs (via reducing scale economies or increasing advertising costs, for example), higher operating costs might be expected under incentive regulation regimes. Fourth, although many telecommunications firms reduced the size of their labour force when they began to operate under incentive regulation, the labour force reductions may have actually increased short term labour costs. This is because labour force reductions are typically achieved by paying employees to terminate their employment voluntarily, and the associated expenditures are often recorded as short term costs (Kridel et al. 1996)

7.3. Network modernisation

As noted above, incentive regulation has been linked to increased network modernisation. Greenstein et al. (1995) report substantial increases in the deployment of fibre optic cable and modern switching equipment under incentive regulation in general and price cap regulation in particular between 1986 and 1991[82]. To illustrate, the authors estimate that price cap regulation leads to a 100 percent increase in the deployment of fibre optic cable relative to rate of return regulation. The corresponding increases under rate case moratoria and earnings sharing regulation are estimated to be 40 percent and 50 percent, respectively. However, if an earnings sharing requirement is appended to a price cap regulation plan, there is virtually no increase in fibre deployment relative to rate of return regulation.

Ai and Sappington (1998) report a less dramatic, but still positive impact of incentive regulation on network modernisation. The authors find that between 1992 and 1996, price cap regulation, earnings sharing regulation, and rate case moratoria

[82] Taylor, Zarkadas, and Zona (1992) provide some corroborating evidence. The authors estimate that incentive regulation accelerated substantially the deployment of modern switching and transmission equipment in the U.S. telecommunications industry in the late 1980s and early 1990s. In contrast, Tardiff and Taylor (1993) find little evidence that incentive regulation accelerates network modernisation.

are all associated with telecommunications networks that contain a larger fraction of modern network switches. Price cap regulation and earnings sharing regulation are also linked to networks with a higher fraction of fibre optic cable[83].

Ai and Sappington (AS)'s smaller estimates of the impact of incentive regulation on network modernisation may stem in part from the longer and more recent time period the authors study. Conceivably, pronounced increases in network modernisation may occur at the onset of incentive regulation, while subsequent increases are less pronounced. It is also possible that particularly large increases in network modernisation were mandated during the early years of incentive regulation. Neither Greenstein et al. nor AS differentiate among incentive regulation plans according to the extent of network modernisation the plans mandate.

The smaller effects of incentive regulation on network modernisation that AS find may also reflect in part the authors' direct measure of competition. AS include as an explanatory variable the number of miles of fibre optic cable that competitive access providers (CAPs) have installed in the state[84]. Greenstein et al. employ alternative measures of competitive pressures, such as whether APs are authorised to operate in the state and whether intraLATA toll competition is permitted in the state[85].

7.4. Total factor productivity

There is mixed evidence regarding the extent to which incentive regulation spurs productivity growth. Productivity is the ratio of a firm's outputs to the inputs it employs in production. Thus, productivity measures the firm's efficiency in transforming inputs into outputs. A firm's total factor productivity growth (TFPG) rate is a measure of the rate at which its productivity is increasing.

Kwoka (1993a) reports a substantial increase in British Telecom's TFPG rate immediately after the onset of privatisation and price cap regulation. He estimates that BT's productivity increased by 21.8 percent between 1984 and 1987, and that

[83] Ai and Sappington also find that aggregate investment is slightly higher under earnings sharing regulation than under rate of return regulation. The authors do not distinguish between earnings sharing plans and price cap regulation plans with an earnings sharing requirement, as Greenstein et al. do.

[84] AS implicitly treat the entire state as the relevant operating territory of the RBOC on which regulation is imposed. Greenstein et al. (1995) employ more precise measures of each RBOC's operating territory.

[85] Wolak (1996b) compares infrastructure modernisation in three U.S. states that have historically fostered industry competition (California, New York, and Illinois) with two states that have tended to be less supportive of competition (Arkansas and Texas). He finds that modernisation proceeded more rapidly in the former states initially, but that initial differences on most dimensions ultimately disappeared. Woroch (2000) finds that investment by competitive local exchange carriers in fibre rings around major cities in the United States spurred corresponding investment by incumbent local exchange carriers between 1983 and 1992.

22 percent of this increase was due to the change in ownership and regulatory structure[86]. Kwoka attributes 52 percent of the increase to the impact of expanded output and scale economies, and the remaining 26 percent of the increase to technical change.

Schmalensee and Rohlfs (1992) estimate that AT&T's higher TFPG under price cap regulation between 1989 and 1991 resulted in extra benefits of more than $1.8 billion for AT&T and its customers. While a significant portion of this sum may be attributable to the effects of price cap regulation, much of it is likely due to the effects of competition[87]. Tardiff and Taylor (1993) estimate that the TFPG rate of large telecommunications firms in the U.S. increased by 2.8 percentage points under incentive regulation prior to 1992. This substantial increase is attributable in roughly equal parts to an increase in the growth rate of outputs and a decrease in the growth rate of inputs under incentive regulation. Of course, if the firms that choose to operate under incentive regulation are those that are particularly capable of increasing their TFPG rates substantially, then the 2.8 percentage point increase likely exceeds the corresponding average increase that would arise if incentive regulation were applied to all firms.

In a more recent study, Resende (1999) finds no evidence that incentive regulation increases productivity growth rates. Resende examines these rates for the major local exchange companies in the U.S. telecommunications industry between 1988 and 1994. The author estimates a translog cost function in order to decompose observed changes in productivity growth rates into effects due to technical change, operating scale, price levels, and incentive regulation. Once the first three effects are controlled for, incentive regulation has no additional impact on the observed productivity growth rate in Resende's sample. The author points out that his sample exhibits considerable heterogeneity and his analysis presumes the firms to operate on the efficiency frontier. Consequently, he suggests that his finding of no overall impact of incentive regulation on productivity growth rates "should be considered with caution" (Resende 1999, p. 42)[88].

[86] In contrast, Kwoka (1993a) estimates that the initial impact of the divestiture of AT&T was to reduce the company's TFPG rate. The positive effects of competition and scale economies, though, outweighed the negative impact of divestiture.

[87] Empirical estimates of the impact of competition on TFPG rates in the telecommunications industry vary. Crandall (1991, p. 69) reports a significantly higher TFPG rate in the U.S. telecommunications industry between 1971 and 1983 than between 1961 and 1970. Staranczak et al. (1994) find no significant relationship between facilities-based competition and TFPG in ten OECD countries between 1983 and 1987.

[88] Dennis Weisman has suggested one reason why the initial productivity gains from incentive regulation might be limited. Firms face legal and institutional constraints as they attempt to reduce their work force in response to increased incentives for cost reduction. These constraints often compel the firms to offer similar incentive retirement packages to all employees. If the most highly-skilled employees have the most attractive employment options at other firms, then they may be particularly likely to accept the retirement package offered by the regulated firm. Consequently, the downsizing may reduce average labour productivity in the regulated firm, *ceteris paribus*.

Majumdar (1997) examines the impact of various incentive regulation plans on a measure of technical efficiency rather than total factor productivity growth. Employing an approach developed by Banker et al. (1984), Majumdar constructs a measure of the relative efficiency of 45 local exchange companies in the United States between 1988 and 1993[89]. The efficiency measure relates to the companies' relative performance in transforming inputs (switches, lines, and employees) into outputs (local, intraLATA toll, and interLATA toll calls), and requires no explicit assumptions about the firms' production technologies. Majumdar finds that: (1) price cap regulation is associated with significant improvement in efficiency after a lag of two years; (2) the positive impact of price cap regulation is diminished when earnings sharing provisions are added to price cap regulation; and (3) earnings sharing regulation alone is associated with a long-term reduction in measured efficiency.

As Majumdar (1997) points out, it is important to determine whether his findings persist in studies that employ a longer time series and more recent data. Alternative measures of efficiency should also be explored. In addition, the endogeneity of the selected regulatory regime should be accounted for explicitly. Better controls for the competitive pressures the firms face and relevant differences in their operating conditions (e.g., customer population density) would also be useful.

7.5. Earnings

There is some evidence that the earnings of the regulated firm increase when it operates under incentive regulation, particularly price cap regulation. The Federal Communications Commission (1992) reports that AT&T's average annual rate of return was 13.2 percent under the first thirty months of price cap regulation, whereas its prescribed rate of return was 12.0 percent just prior to the imposition of price cap regulation in July 1 1989. Tardiff and Taylor (1993) find no significant relationship between earnings and incentive regulation, broadly defined, between 1984 and 1990.

Armstrong et al. (1994, pp. 204–205) report that British Telecom's rate of return on capital increased from less than 17 percent in 1984 (when price cap regulation was instituted) to more than 21 percent in 1991 and 1992. Of course, this increase reflects a variety of effects, including the change from public to private ownership of BT. Ai and Sappington (1998) report that RBOC earnings were higher (by approximately 16 percent) under price cap regulation than under rate of return regulation between 1991 and 1996. However, they do not find evidence of higher earnings under other forms of incentive regulation. Of course, studies of the effects

[89] Majumdar (1995), Jha and Majumdar (1999), Fraquelli and Vannoni (2000), and Resende (2000) employ related methodologies and find that competition and/or incentive regulation have motivated suppliers of telecommunications services to increase their operating efficiencies in recent years.

of regulatory policy on earnings are difficult to conduct because firms often have significant discretion in allocating revenues and costs across time periods.

7.6 Service quality

Incentive regulation can promise the regulated firm substantial financial reward for realised cost reductions. Consequently, the possibility arises that a firm that operates under incentive regulation may be tempted to reduce service quality unduly in order to reduce operating costs[90]. The evidence on this issue is mixed.

AT&T experienced some large-scale service outages while operating under price cap regulation in 1990 and 1991. However, after careful study of the circumstances surrounding the outages, the U.S. Federal Communications Commission (1992) reported no evidence of a link between price cap regulation and AT&T's service quality. Tardiff and Taylor (1993) also find no evidence that incentive regulation promotes reduced service quality in the U.S. telecommunications industry.

Armstrong et al. (1994) report widespread dissatisfaction with the service quality provided by British Telecom (BT) during the early years of price cap regulation. An audit by the regulatory body Oftel revealed that although BT's service quality was not as poor as it was generally perceived to be, it could have been substantially higher than it was on many dimensions. BT was required to publish certain service quality performance statistics twice each year, and was also subjected to financial penalties for poor service quality. Increased competition was also encouraged, and appeared to have had its intended effect. To illustrate, 17 percent of BT's pay telephones were out of order on average in 1986. The corresponding statistic in 1987 was 23 percent. In 1987, BT's primary competitor, Mercury, was authorised to provide pay telephone service. By 1989, less than 5 percent of BT's pay telephones were out of order (Rovizzi and Thompson, 1992).

Ai and Sappington (1998) find that during the mid 1990's, the service quality provided by the RBOCs in the United States was lower on some dimensions under incentive regulation than under rate of return regulation. Most notably, the RBOCs remedied reported service problems more slowly under price cap regulation, earnings sharing regulation, and rate case moratoria than they did under rate of return regulation. Residential customers also reported more service problems (per access line) under earnings sharing regulation than under rate of return regulation.

But Ai and Sappington (1998) also report higher service quality on some dimensions under incentive regulation. For example, residential and business customers both registered fewer complaints with their public utility commission under price cap regulation, earnings sharing regulation, and rate case moratoria than under rate of return regulation. Residential and business customers also received more timely installation of new telephone service under earnings sharing

[90] Norsworthy and Tsai (1999) stress the need to account for realised service quality when designing reward structures for regulated firms.

regulation. In addition, commitments to install new telephone service for business customers were met more frequently under rate case moratoria. Improved service quality may arise under rate case moratoria in part because of demonstration effects. Rate case moratoria are often implemented as temporary regulatory regimes while the details of alternative regimes are negotiated. Consequently, when it operates under a rate case moratorium, a regulated firm may take special precautions against service quality deterioration, to ensure that support for an attractive future regulatory regime is not eroded.

7.7. Universal service

The evidence to date suggests that certain forms of incentive regulation may increase telephone penetration rates (i.e., the fraction of households with a telephone), at least in the United States. Tardiff and Taylor (1993) find that the telephone penetration rate is approximately 1 percentage point higher in states where earnings sharing regulation is imposed than in states where rate of return regulation is employed, after controlling for other relevant factors. Ai and Sappington (1998) find a similar increase under rate case moratoria relative to rate of return regulation. Conceivably, the lower prices that incentive regulation plans induce or mandate could promote increased telephone penetration rates. Regulated firms might also undertake special efforts under incentive regulation to promote universal service in order to garner popular support for favourable future regulatory regimes. This effect may be most pronounced under rate case moratoria, since these regimes often serve as transitions to alternative forms of incentive regulation, such as earnings sharing or price cap regimes.

It is particularly important to control for the effects of competition when assessing the impact of regulation on telephone penetration rates. Although competition can undo cross subsidies that have been implemented to promote access to the telephone network (e.g., pricing basic local service below cost), it can also reduce the average level of service prices and thereby increase the net gains that consumers perceive from securing access to the network. Consequently, the impact of competition on universal service is ambiguous, *a priori*. The works of Hausman et al. (1993) and Wolak (1996a) suggest that increased competition is associated with increased telephone penetration rates in the United States. Barros and Seabra (1999) report mixed evidence in other OECD countries. None of these studies examines the effect of the regulatory regime on universal service rates[91].

[91] Eriksson et al. (1998) report that targeted subsidies based on financial need have been a more cost-effective means of increasing telephone subscribership in the U.S. than have corresponding untargeted subsidies. Cain and MacDonald (1991) find that moderate increases in the price of unmeasured local service have limited impact on telephone subscription rates in the U.S. when the rate for measured local service is low. Neither of these studies measure the impact of incentive regulation on telephone subscription.

7.8. Summary

In summary, the studies to date provide varied evidence regarding the impact of incentive regulation on performance in the U.S. telecommunications industry. Incentive regulation appears to increase the deployment of modern switching and transmission equipment, to spur an increase in total factor productivity growth, and to foster a modest reduction in certain service prices. There is little evidence, though, that incentive regulation leads to a significant reduction in operating costs. There is some evidence that earnings may be higher under price cap regulation. There is little evidence of a systematic decline in service quality under incentive regulation.

Overall, incentive regulation appears to affect industry performance, but the effects are generally not dramatic. This may be the case because the key differences among the various forms of incentive regulation in practice are often less pronounced than their different classifications on paper might suggest. For instance, as noted in Section 4, if prices are re-set at the end of each price cap period to eliminate any extra-normal profit that the firm may have generated during the prevailing phase of the price cap regime, then price cap regulation may function much like rate of return regulation with a specified regulatory lag[92]. Knowing that exceptional current performance will call forth more exacting future standards, the regulated firm may rationally choose not to operate at peak efficiency, even though it might receive a short-term financial reward for doing so.

8. Conclusions

The telecommunications industry has experienced and continues to experience a shift from rate of return regulation to a variety of forms of incentive regulation, including price cap regulation. By focusing more on the control of prices than the control of earnings, incentive regulation plans can provide stronger incentives for the regulated firm to reduce its production costs and increase its operating revenues. By affording the firm greater flexibility in setting prices, incentive regulation plans can also empower incumbent suppliers to meet emerging competitive challenges more effectively.

Of course, no regulatory plan is a perfect substitute for the discipline of competitive markets. Therefore, although price cap regulation, for example, can offer some advantages over rate of return regulation, price cap regulation is not without its own potential drawbacks. For instance, it can facilitate extreme earnings for the firm, it may provide inadequate incentives for service quality, and it can invite strategic pricing to relax the price cap constraint. Ever aware of these potential dangers, regulators often modify price cap plans in ways that render

[92] See Baron (1991), Pint (1992), and Liston (1993) for additional discussion of this observation.

them less distinct from rate of return regulation. Common modifications include earnings sharing requirements, revenue sharing requirements, Z factors, and short time spans between reviews that serve to match anticipated revenues to realised operating costs.

In part because of these modifications and the resulting difficulty in measuring all relevant differences among regulatory plans, the estimated impact of incentive regulation on performance in telecommunications markets has generally not been dramatic. There is some evidence of higher earnings, more rapid network modernisation, and lower prices for certain services under incentive regulation, but there is little evidence of significant change in operating costs, service quality, aggregate investment, or telephone penetration rates under incentive regulation.

Of course, incentive regulation is a relatively recent phenomenon in the telecommunications industry. Therefore, current estimates of the impact of incentive regulation may not reflect accurately its true long term impact. More accurate estimates must await the arrival of a longer time series of data. More comprehensive data sets will also admit finer classifications of incentive regulation plans. Finer classifications will facilitate a better understanding of the impact of individual elements of incentive regulation plans. Systematic identification of the environments in which particular regulatory plans have the most pronounced and the most favourable impact on industry performance awaits future research.

Future empirical work must distinguish more clearly between the effects of incentive regulation and the effects of competition. Additional theoretical work on the impact of competition would also be valuable. Most analyses in the literature take as given the set of regulated products and services that are regulated. In practice, one of the most difficult tasks that regulators face in today's telecommunications industry is determining when it is appropriate to substitute market discipline (and anti-trust scrutiny) for regulatory control (Barnich, 1992). Simple, practical rules to govern this decision would be useful. Simple rules for adjusting regulatory policy as competition emerges would also be valuable[93].

It is also important to recognise that the extent of market competition in a particular telecommunications industry is seldom entirely exogenous. Regulatory rules typically play a central role in determining the extent and nature of competition[94]. The optimal design of incentive regulation in the presence of both actual and potential competition is an issue that warrants more attention in the coming years.

[93] In order to tailor regulatory policy to the extent of market competition, regulators must be able to assess accurately the degree of prevailing market competition. On-going extensive data collection by regulatory agencies is likely to prove valuable in this regard (Bernstein et al., 1996).

[94] See Laffont and Tirole (1993; 1999) for useful reviews of the literature on the optimal design of competition for the right to operate in a specified market.

References

Abel, J., 2000, The performance of the state telecommunications industry under price-cap regulation: An assessment of the empirical evidence, National Regulatory Research Institute Report #00–14.

Abel, J., 1999, Dominant firm pricing with price-cap regulation and fringe competition: An economic analysis of local telephone pricing, Mimeo, NRRI/Ohio State University.

Acton, J. and I. Vogelsang, 1989, Introduction to the symposium on price cap regulation, RAND Journal of Economics, 20, 369–372.

Ai, C. and D. Sappington, 1998, The impacts of state incentive regulation on the U.S. telecommunications industry, University of Florida Discussion Paper.

Alexander, I. and T. Irwin, 1996, Price caps, rate-of-return regulation, and the cost of capital, Private Sector, Note No. 87, The World Bank.

Armstrong, M., 2002, The theory of access pricing and interconnection, in M. E. Cave, S. K. Majumdar, and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science B. V., 295–384.

Armstrong, M., S. Cowan and J. Vickers, 1994, Regulatory reform: Economic analysis and British experience. Cambridge, MA: MIT Press.

Armstrong, M., S. Cowan and J. Vickers, 1995, Non-linear pricing and price cap regulation, Journal of Public Economics, 58, 33–55.

Armstrong, M., R. Rees and J. Vickers, 1995, Optimal regulatory lag under price cap regulation, Revista Espanola De Economia 10, 93–116.

Armstrong, M. and D. Sappington, 2002, Recent developments in the theory of regulation, in M. Armstrong and R. Porter, eds., Handbook of Industrial Organisation, Amsterdam, Elsevier Science B. V.

Armstrong, M. and J. Vickers, 1991, Welfare effects of price discrimination by a regulated monopolist, RAND Journal of Economics, 22, 571–580.

Armstrong, M. and J. Vickers, 1993, Price discrimination, competition and regulation, Journal of Industrial Economics, 41, 335–360.

Averch, H. and L. Johnson, 1962, Behaviour of the firm under regulatory constraint, American Economic Review, 52, 1053–1069.

Bailey, E. and R. Coleman, 1971, The effect of lagged regulation in an Averch-Johnson model, Bell Journal of Economics, 2, 278–292.

Banker, R. D., H. H. Chang and S. K. Majumdar, 1995, The consequences of evolving competition on components of firms profits: Recent evidence from the U.S. telecommunications industry, Information Economics and Policy, 7, 37–56.

Banker, R. D., H. H. Chang and S. K. Majumdar, 1996, Profitability, productivity and price recovery patterns in the U.S. telecommunications industry, Review of Industrial Organisation, 11, 1–17.

Banker, R., A. Charnes and W. Cooper, 1984, Some models for estimating technical and scale efficiencies in data envelopment analysis, Management Science, 30, 1078–1092.

Barnich, T., 1992, The challenge for incentive regulation, Public Utilities Fortnightly, 129, 15–17.

Baron, D., 1989, Design of regulatory mechanisms and institutions, in R. Schmalensee and R. Willig, eds., Handbook of Industrial Organisation: Volume II. Amsterdam: North-Holland, 1347–1447.

Baron, D., 1991, Information, incentives, and commitment in regulatory mechanisms: Regulatory innovation in telecommunications, in M. Einhorn, ed., Price Caps and Incentive Regulation in Telecommunications. Boston: Kluwer Academic Publishers, 47–75.

Baron, D. and R. Myerson, 1982, Regulating a monopolist with unknown costs, Econometrica, 50, 911–930.

Barros, P., and M. Seabra, 1999, Universal service: Does competition help or hurt? Information Economics and Policy, 11, 45–60.

Baumol, W. and A. Klevorick, 1970, Input choices and rate-of-return regulation: An overview of the discussion, Bell Journal of Economics, 1, 162–190.

Beesley, M. and S. Littlechild, 1989, The regulation of privatised monopolies in the United Kingdom, RAND Journal of Economics, 20, 454–472.
BellSouth Services Regulatory Matters, 1987–1992, Regulatory reform: A nation-wide summary, edited by A. Helms. Vols. 2–12. Atlanta, GA.
BellSouth Telecommunications Regulatory Policy and Planning, 1993, Regulatory reform: A nation-wide summary, edited by B. Harber and L. Greer. Vols. 13 and 14. Atlanta, GA.
BellSouth Telecommunications – State Regulatory Matters, 1994–1995, Regulatory reform: A nation-wide summary. Vols. 15–17. Atlanta, GA.
Berg, S. and J. Lynch, 1992, The measurement and encouragement of telephone service quality, Telecommunications Policy, 16, 210–224.
Bernstein, J., K. Gordon, S. Levin, D. Sappington and D. Weisman, 1996, A price cap regulation plan for the Canadian telecommunications industry, AGT Ltd. Report.
Bernstein, J. and D. Sappington, 1999a, Setting the *X* factor in price cap regulation plans, Journal of Regulatory Economics, 16, 5–25.
Bernstein, J. and D. Sappington, 2000, How to determine the X in RPI – X regulation: A users guide, Telecommunications Policy, 24, 63–68.
Besanko, D. and D. Sappington, 1987, Designing regulatory policy with limited information. London: Harwood Academic Publishers.
Biglaiser, G. and M. Riordan, 2001, Dynamics of price regulation, RAND Journal of Economics, forthcoming.
Blackmon, G., 1994, Incentive regulation and the regulation of incentives. Boston: Kluwer Academic Publishers.
Blank, L., D. Kaserman and J. Mayo, 1998, Dominant firm pricing with competitive entry and regulation: The case of intraLATA toll, Journal of Regulatory Economics, 14, 35–54.
Bradley, I. and C. Price, 1988, The economic regulation of private industries by price constraints, Journal of Industrial Economics, 37, 99–106.
Braeutigam, R., M. Magura, and J. Panzar, 1997, The effects of incentive regulation on U.S. local telephone rates, Mimeo, Northwestern University.
Braeutigam, R. and J. Panzar, 1989, Diversification incentives under price-based and cost-based regulation, RAND Journal of Economics, 20, 373–391.
Braeutigam, R. and J. Panzar, 1993, Effects of the change from rate-of-return to price-cap regulation, American Economic Review, 83, 194–198.
Brennan, T., 1989, Regulating by 'capping prices,' Journal of Regulatory Economics, 1, 133–147.
Brennan, T., 1990, Cross-subsidisation and cost misallocation by regulated monopolists, Journal of Regulatory Economics, 2, 37–52.
Brennan, T. and K. Palmer, 1994, Comparing the costs and benefits of diversification by regulated firms, Journal of Regulatory Economics, 6, 115–136.
Brock, G., 1994, Telecommunications policy for the information age. Cambridge, MA: Harvard University Press.
Cabral, L. and M. Riordan, 1989, Incentives for cost reduction under price cap regulation, Journal of Regulatory Economics, 1, 93–102.
Caillaud, B., R. Guesnerie, P. Rey and J. Tirole, 1988, Government intervention in production and incentives theory: A review of recent contributions, RAND Journal of Economics 19, 1–26.
Cain, P. and J. Macdonald, 1991, Telephone pricing structures: The effects on universal service, Journal of Regulatory Economics, 3, 293–308.
Cave, M., 2000, Price-capping mechanisms, in T. Jenkinson, ed. Readings in Microeconomics, Oxford: Oxford University Press, 224–238, Second Edition.
Chang, Y and J. Warren, 1997, Allocative efficiency and diversification under price-cap regulation, Information Economics and Policy, 9, 3–17.
Christensen, L., P. Schoech and M. Meitzen, 1994, Productivity of the local operating companies subject to price cap regulation, Mimeo, Christensen Associates.

Clemenz, G., 1991, Optimal price cap regulation, Journal of Industrial Economics, 39, 391–408.
Cowan, S., 1997a, Price-cap regulation and inefficiency in relative pricing, Journal of Regulatory Economics, 12, 53–70.
Cowan, S., 1997b, Tight average revenue regulation can be worse than no regulation, Journal of Industrial Economics, 45, 75–88.
Crandall, R., 1991, After the break-up: U.S. telecommunications in a more competitive era, Washington, D.C.: The Brookings Institution.
Crandall, R. and L. Waverman, 1995, Talk is cheap: The promise of regulatory reform in North American telecommunications, Washington, D.C.: The Brookings Institution.
Crew, M. and K. Crocker, 1991, Diversification and regulated monopoly, in M. Crew, ed., Competition and the Regulation of Utilities. Boston: Kluwer Academic Publishers, 33–49.
Crew, M. and P. Kleindorfer, 1996a, Incentive regulation in the United Kingdom and the United States: Some lessons, Journal of Regulatory Economics, 9, 211–225.
Crew, M. and P. Kleindorfer, 1996b, Price caps and revenue caps: Incentives and disincentives for efficiency, in M. Crew ed., Pricing and Regulatory Innovations Under Increasing Competition, Boston: Kluwer Academic Publishers, 39–52.
Cubbin, J. and G. Tzanidakis, 1998, Regression versus data envelopment analysis for efficiency measurement: An application to the England and Wales regulated water industry, Utilities Policy, 7, 75–85.
Davis, R., 2000, Acting on performance-based regulation, The Electricity Journal, 13, 13–15.
Diewert, W. E. and A. Nakamura, 1999, Benchmarking and the measurement of best practice efficiency: An electricity generation application, Canadian Journal of Economics, 32, 570–588.
Donald, S. and D. Sappington, 1995, Explaining the choice among regulatory plans in the U.S. telecommunications industry, Journal of Economics and Management Strategy, 4, 237–265.
Donald, S. and D. Sappington, 1997, Choosing among regulatory options in the United States telecommunications industry, Journal of Regulatory Economics, 12, 227–243.
Encinosa, W. and D. Sappington, 1995, Toward a benchmark for optimal prudency policy, Journal of Regulatory Economics, 7, 111–131.
Eriksson, R., D. Kaserman and J. Mayo, 1998, Targeted and untargeted subsidy schemes: Evidence from post-divestiture efforts to promote universal telephone service, Journal of Law and Economics, 41, 477–502.
Federal Communications Commission, 1992, Price cap performance review for AT&T, CC Docket No. 92–134, Notice of Inquiry.
Federal Communications Commission, 1999, Access charge reform, Fifth report and order and further notice of proposed rulemaking, in CC Docket No. 96–262, Adopted.
Foreman, D., 1995, Pricing incentives under price-cap regulation, Information Economics and Policy, 7, 331–351.
Fraquelli, G. and D. Vannoni, 2000, Multidimensional performance in telecommunications, regulation, and competition: Analysing the European major players, Information Economics and Policy, 12, 27–46.
Fraser, R., 1995, The relationship between the costs and the prices of a multi-product monopoly: The role of price-cap regulation, Journal of Regulatory Economics, 8, 23–32.
Galal, A., 1996, Chile: Regulatory specificity, credibility of commitment, and distributional demands, in B. Levy and P. T. Spiller, eds., Regulations, Institutions, and Commitment: Comparative Studies of Telecommunications, Cambridge: Cambridge University Press.
Gasmi, F., M. Ivaldi, and J.-J Laffont, 1994, Rent extraction and incentives for efficiency in recent regulatory proposals, Journal of Regulatory Economics, 6, 151–176.
Giulietti, M and C. Waddams, 2000, Price, incentive regulation and efficient pricing: Empirical evidence, University of Warwick Discussion Paper No. 00/2.
Green, R. and M. R. Pardina, 1999, Resetting price controls for privatised utilities: A manual for regulators. Washington, D.C.: The World Bank.

Greenstein, S., S. McMaster and P. Spiller, 1995, The effect of incentive regulation on infrastructure modernisation: Local exchange companies' deployment of digital technology, Journal of Economics and Management Strategy, 4, 187–236.
Hausman, J., Tardiff, T. and A. Belinfante, 1993, The effects of the break-up of AT&T on telephone penetration in the United States, American Economic Review, 83, 178–184.
Hillman, J. and R. Braeutigam, 1989, Price level regulation for diversified public utilities. Norwell, MA: Kluwer Academic Press.
Howe, S., 1983, A survey of regulatory treatment of plant cancellation costs, Public Utilities Fortnightly, 111, 52–58.
Howe, S., 1985, The used and useful standard: What should the criteria be? Public Utilities Fortnightly, 115, 54–56.
Hunt, L. and E. Lynk, 1995, Privatisation and efficiency in the U.K. water industry: An empirical analysis, Oxford Bulletin of Economics and Statistics, 57, 371–388.
Isaac, R. M., 1991, Price cap regulation: A case study of some pitfalls of implementation, Journal of Regulatory Economics, 3, 193–210.
Jha, R. and S. Majumdar, 1999, A matter of connections: OECD telecommunications sector productivity and the role of cellular technology diffusion, Information Economics and Policy, 11, 243–269.
Joskow, P.,1974, Inflation and environmental concern: Structural change in the process of public utility price regulation, Journal of Law and Economics, 17, 291–327.
Joskow, P. and R. Schmalensee, 1986, Incentive regulation for electric utilities, Yale Journal on Regulation, 4, 1–50.
Kaestner, R. and B. Kahn, 1990, The effects of regulation and competition on the price of AT&T intrastate telephone service, Journal of Regulatory Economics, 2, 363–377.
Kang, J, D. Weisman, and M. Zhang, 2000, Do consumers benefit from tighter price cap regulation? Economics Letters, 67, 113–119.
Kirchhoff, H., 1994a, Telcos in East examine price caps to tackle competition, State Telephone Regulation Report, 12-3, 1–7.
Kirchhoff, H., 1994b, Alternate regulation seeks foothold in rocky West, State Telephone Regulation Report, 12-4, 1–7.
Kirchhoff, H., 1995a, RHCs in East convince 8 states to embrace price cap regulation, State Telephone Regulation Report,13-3, 5–9.
Kirchhoff, H., 1995b, Western states may see major swing to price caps this year, State Telephone Regulation Report, 13-4, 6–11.
Kirchhoff, H., 1996a, Two-thirds of Eastern states have switched to price caps, State Telephone Regulation Report, 14-2, 10–14.
Kirchhoff, H., 1996b, Price caps still struggling for acceptance in Western states, State Telephone Regulation Report, 14-3, 8–13.
Kirchhoff, H., 1997a, Price cap regulation has become the norm in the East, State Telephone Regulation Report, 15-6, 1–9.
Kirchhoff, H., 1997b, Price caps struggle in Western states, But '97 may see major changes, State Telephone Regulation Report, 15-7, 1–8.
Kirchhoff, H., 1998a, Earnings regulation for big incumbent telcos just about extinct in Eastern U.S., State Telephone Regulation Report, 16-7, 1–4.
Kirchhoff, H., 1998b, Price caps still struggle in Western states, But '98 may see some changes, State Telephone Regulation Report, 16-8, 1–2.
Kirchhoff, H., 1998a, Price caps are 'standard' regulatory method for Eastern U.S. incumbent telcos, State Telephone Regulation Report, 17-17, 1–3.
Kirchhoff, H., 1999b, Price caps take solid hold in Western U.S., rate-of-return lingers, State Telephone Regulation Report 17-18, 1–3.
Knittel, C., 1997, Local telephone pricing: Competitive forces at work, Utilities Policy, 6, 87–96.

Kolbe, L. and W. Tye, 1991, The Duquesne opinion: How much 'hope' is there for investors in regulated firms? Yale Journal on Regulation, 8, 113–157.
Kridel, D., D. Sappington and D. Weisman, 1996, The effects of incentive regulation in the telecommunications industry: A survey, Journal of Regulatory Economics, 9, 269–306.
Kwoka, J., 1991, Productivity and price caps in telecommunications, in M. Einhorn, ed., Price Caps and Incentive Regulation in Telecommunications, Boston: Kluwer Academic Publishers, 77–93.
Kwoka, J., 1993a, The effects of divestiture, privatisation, and competition on productivity in U.S. and U. K. telecommunications, Review of Industrial Organisation, 8, 49–61.
Kwoka, J., 1993b, Implementing price caps in telecommunications, Journal of Policy Analysis and Management, 12, 726–752.
Laffont, J.-J, 1994, The new economics of regulation ten years after, Econometrica, 62, 507–538.
Laffont, J.-J and J. Tirole, 1986, Using cost observation to regulate firms, Journal of Political Economy, 94, 614–641.
Laffont, J.-J and J. Tirole, 1993, A theory of incentives in procurement and regulation, Cambridge, MA: The MIT Press.
Laffont, J.-J and J. Tirole, 1996, Creating competition through interconnection: Theory and practice, Journal of Regulatory Economics, 10, 227–256.
Laffont, J.-J and J. Tirole, 1999, Competition in telecommunications, Cambridge, MA: The MIT Press.
Law, P., 1995, Tighter average revenue regulation can reduce consumer welfare, Journal of Industrial Economics, 42, 399–404.
Law, P., 1997, Welfare effects of pricing in anticipation of Laspeyres price cap regulation: An example, Bulletin of Economic Research, 49, 17–27.
Lehman, D. and D. Weisman, 2000, The political economy of price cap regulation, Review of Industrial Organisation, 16, 343–356.
Levy, B. and P. Spiller, 1994, The institutional foundations of regulatory commitment: A comparative analysis of telecommunications, Journal of Law, Economics, and Organisation, 10, 201–246.
Levy, B. and P. Spiller, 1996, Regulations, institutions, and commitment: Comparative studies of telecommunications, Cambridge: Cambridge University Press.
Lewis, T. and D. Sappington, 1989, Regulatory options and price cap regulation, RAND Journal of Economics, 20, 405–416.
Liston, C., 1993, Price-cap versus rate-of-return regulation, Journal of Regulatory Economics, 5, 25–48.
Lynch, J., T. Buzas, and S. Berg, 1994, Regulatory measurement and evaluation of telephone service quality, Management Science, 40, 169–194.
Lyon, T., 1991, Regulation with 20-20 hindsight: 'Heads I win, tails you lose,' RAND Journal of Economics, 22, 581–595.
Lyon, T., 1992, Regulation with 20-20 hindsight: Least-cost rules and variable costs, Journal of Industrial Economics, 40, 277–289.
Lyon, T., 1995, Regulatory hindsight review and innovation by electric utilities, Journal of Regulatory Economics, 7, 233–254.
Lyon, T., 1996, A model of sliding-scale regulation, Journal of Regulatory Economics, 9, 227–247.
Magura, M., 1998, Incentive regulation and local exchange carrier pricing policies, Northwestern University Discussion Paper,
Majumdar, S. K., 1995, X-efficiency in emerging competitive markets: The case of U. S. telecommunications, Journal of Economic Behaviour and Organisation, 26, 129–144.
Majumdar, S. K., 1997, Incentive regulation and productive efficiency in the U.S. telecommunications industry, Journal of Business, 70, 547–576.
Mansell, R. and J. Church, 1995, Traditional and incentive regulation: Applications to natural gas pipelines in Canada, Calgary: The Van Horne Institute.
Mathios, A. and R. Rogers, 1989, The impact of alternative forms of state regulation of AT&T on direct-dial, long-distance telephone rates, RAND Journal of Economics, 20, 437–53.

Mayer, C. and J. Vickers, 1996, Profit-sharing regulation: An economic appraisal, Fiscal Studies, 17, 1–18.
Mitchell, B. and I. Vogelsang, 1991, Telecommunications pricing: Theory and practice, Cambridge: Cambridge University Press.
Mueller, M., 1993, Telephone companies in paradise: A case study in telecommunications deregulation, New Brunswick, NJ: Transaction Publishers.
Neri, J. and K. Bernard, 1994, Price caps and rate of return: The British experience, Public Utilities Fortnightly, 132, 34–36.
Neu, Werner, 1993, Allocative inefficiency properties of price-cap regulation, Journal of Regulatory Economics, 5, 159–182.
Newbery, D., 1999, Privatisation, restructuring, and regulation of network utilities, Cambridge, MA: The MIT Press.
Norsworthy, J. and D. Tsai, 1999, The role of service quality and capital technology in telecommunications regulation, Information Economics and Policy, 11, 127–146.
Office of Telecommunications (Oftel), 2000, Price Control Review, London.
OECD, 1999, Communications Outlook, Paris.
Palmer, K., 1991, Diversification by regulated monopolies and incentives for cost-reducing R&D, American Economic Review, 81, 266–270.
Pint, E., 1992, Price-cap versus rate-of-return regulation in a stochastic-cost model, RAND Journal of Economics, 23, 564–578.
Ramsey, F., 1927, A contribution to the theory of taxation, Economic Journal, 37, 47–61.
Resende, M., 1999, Productivity growth and regulation in U.S. local telephony, Information Economics and Policy, 11, 23–44.
Resende, M., 2000, Regulatory regimes and efficiency in U.S. local telephony, Oxford Economics Papers, 52, 447–470.
Rovizzi, L. and D. Thompson, 1992, The regulation of product quality in the public utilities and the citizens charter, Fiscal Studies, 13, 74–95.
Sappington, D., 1980, Strategic firm behaviour under a dynamic regulatory adjustment process, Bell Journal of Economics, 11, 360–372.
Sappington, D., 1994, Designing incentive regulation, Review of Industrial Organisation, 9, 245–272.
Sappington, D. and D. Sibley, 1992, Strategic non-linear pricing under price cap regulation, RAND Journal of Economics, 23, 1–19.
Sappington, D. and J. Stiglitz, 1987, Information and regulation, in E. Bailey, ed., Public Regulation: New Perspectives on Institutions and Policies, Cambridge, MA: MIT Press.
Sappington, D. and D. Weisman, 1994a, Designing superior incentive regulation: Accounting for all of the incentives all of the time, Public Utilities Fortnightly, 132, 12–15.
Sappington, D. and D. Weisman, 1994b, Designing superior incentive regulation: modifying plans to preclude recontracting and promote performance, Public Utilities Fortnightly, 132, 27–32.
Sappington, D. and D. Weisman, 1996a, Designing incentive regulation for the telecommunications industry, Cambridge, MA: MIT Press.
Sappington, D. and D. Weisman, 1996b, Potential pitfalls in empirical investigations of the effects of incentive regulation plans in the telecommunications industry, Information Economics and Policy, 8, 125–140.
Sappington, D. and D. Weisman, 1996c, Revenue sharing in incentive regulation plans, Information Economics and Policy, 8, 229–248.
Sawkins, J., 1995, Yardstick competition in the English and Welsh water industry, Utilities Policy, 4, 27–36.
Schmalensee, R., 1989, Good regulatory regimes, RAND Journal of Economics, 20, 417–436.
Schmalensee, R. and J. Rohlfs, 1992, Productivity gains resulting from interstate price caps for AT&T, Mimeo, National Economic Research Associates.
Shin, R. and J. Ying, 1993, Efficiency in regulatory regimes: Evidence from price caps, Presented at the Twenty-First Annual Telecommunications Policy Research Conference: Solomons, MD.

Shleifer, Andrei, 1985, A theory of yardstick competition, RAND Journal of Economics, 16, 319–327.
Sibley, D., 1989, Asymmetric information, incentives, and price cap regulation, RAND Journal of Economics, 20, 392–404.
Sobel, J., 1999, A re-examination of yardstick regulation, Journal of Economics and Management Strategy, 8, 33–60.
Staranczak, G., E. Sepulveda, P. Dilworth and S. Shaikh, 1994, Industry structure, productivity, and international competitiveness: The case of telecommunications, Information Economics and Policy, 6, 121–142.
State of New York Public Service Commission, 1994, Performance regulation plan for New York telephone company, Case 92-C-0665.
State Telephone Regulation Report, 2000a, N.E. states use price cap regulation for big telcos, rate of return for small, minimal for CLECs, 18–20, 1–4.
State Telephone Regulation Report, 2000b, SE and Great Lakes state regulators embrace price caps for their big incumbents, 18–21, 1–3.
State Telephone Regulation Report, 2000c, Western states swing to price caps, but rate-of-return lingers, 18–22, 1–3.
Tardiff, T. and W. Taylor, 1993, telephone company performance under alternative forms of regulation in the U.S., Mimeo, National Economic Research Associates.
Tardiff, T. and W. Taylor, 1996, Revising price caps: The next generation of incentive regulation plans, in M. Crew, ed., Pricing and Regulatory Innovation Under Increasing Competition, Boston: Kluwer Academic Publishers, 21–38.
Taylor, W., C. Zarkadas and J. D. Zona, 1992, Incentive regulation and the diffusion of new technology in telecommunications, Mimeo, National Economic Research Associates.
Train, K., 1991, Optimal regulation: The economic theory of natural monopoly, Cambridge, MA: The MIT Press.
U. S. Department of Commerce, 1991, The NTIA infrastructure report: Telecommunications in the age of information, NTIA Special Publication 91–26, Washington, D.C.
Vickers, J., 1997, Regulation, competition, and the structure of prices, Oxford Review of Economic Policy, 13, 15–26.
Vickers, J. and G. Yarrow, 1988, Privatisation: An economic analysis, Cambridge, MA: The MIT Press.
Vogelsang, I., 1989, Price cap regulation of telecommunications services: A long-run approach, in M. Crew, ed., Deregulation and Diversification of Utilities, Boston: Kluwer Academic Publishers.
Vogelsang, I. and J. Finsinger, 1979, A regulatory adjustment process for optimal pricing by multi-product monopoly firms, Bell Journal of Economics, 10, 151–171.
Waterson, M., 1992, A comparative analysis of methods for regulating public utilities, Metroeconomica, 43, 205–226.
Weisman, D., 1993, Superior regulatory regimes in theory and practice, Journal of Regulatory Economics, 5, 355–366.
Weisman, D., 1994, Why more may be less under price-cap regulation, Journal of Regulatory Economics, 6, 339–362.
Wolak, F., 1996a, Can universal service survive in a competitive telecommunications environment? Evidence from the United States consumer expenditure survey, Information Economics and Policy, 8, 163–203.
Wolak, F., 1996b, Competition and RBOC revenues and infrastructure investment: The case of Arkansas, California, Texas, New York, and Illinois, Mimeo, Stanford University.
Wolfstetter, E., 1999, Topics in microeconomics: Industrial organisation, auctions, and incentives, Cambridge: Cambridge University Press.
Woroch, Glenn, 2000, Competition's effect on investment in digital infrastructure, University of California, Berkeley, Discussion Paper.

This Page Intentionally Left Blank

Chapter 8

THE THEORY OF ACCESS PRICING AND INTERCONNECTION

MARK ARMSTRONG*

Nuffield College, Oxford

Contents

* The section on one-way access pricing owes a great debt to joint work with John Vickers. I am also grateful to the editors, as well as to Eric Bond, Carli Coetzee, Wouter Dessein, Michael Doane, Joshua Gans, Martin Peitz, Patrick Rey, David Sappington, Vasiliki Skreta, Daniel Spulber, Jean Tirole, Tommaso Valletti and Julian Wright for several helpful comments. I am responsible for all views and remaining errors.

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

This chapter discusses the interaction between competition and regulation in telecommunications markets, with the focus being on the central issue of access charges and network interconnection[2]. The main discussion is divided into three parts. In Section 2 I discuss the problem of "one way" access pricing, which is where entrants need to purchase vital inputs from the incumbent, but not *vice versa*. In this case, because of the incumbent's monopoly position in the access market the firm behaves in many ways like a text-book monopolist, and without control it will usually set access charges too high. The desirability of regulatory intervention is therefore clear, and the section discusses how to set access charges in a variety of situations: when the incumbent's retail is chosen in advance (and perhaps in an *ad hoc* way); when the incumbent's retail prices and access charges are chosen simultaneously to maximize welfare (Ramsey pricing), and when the incumbent is free to choose its retail tariff.

In Section 4 I discuss the "two way" access, or network interconnection, problem, which is where all firms in the market need to purchase vital inputs—namely, access to rival firms' subscribers—from each other. In this situation the danger might not be so much one of *foreclosure*, as in the one way case, but of *collusion* between networks. Can free negotiations between networks over their mutual access charges induce high prices for subscribers? The answer to this question is subtle, and depends in part on the kinds of tariffs that networks offer. Bridging these two sections is a discussion in Section 3 of "competitive bottlenecks", which is where networks operate in a competitive market *for* subscribers, and yet have a monopoly position for providing access *to* these subscribers. (This model is motivated by the mobile telephony market and the internet market.) Not surprisingly, this kind of market exhibits features both of competitive markets (no excess profits overall) and of monopoly markets (in the mobile sector, high profits are generated in the market for call termination, which are then used to attract subscribers).

Before starting the analysis, however, it is worth listing a few kinds of access services. There are numerous kinds of entry strategies a new telecommunications firm could follow—ranging from utilizing basic resale of an incumbent firm's own retail service, which involves little in the way of physical investment, to full-blown infrastructure-based entry—and each strategy requires its own kinds of access or interconnection services from incumbent firms. A very partial list of such services includes:

- Call termination: When a customer connected to firm A calls a customer of firm B, the former must (usually) pay the latter to deliver its call. Included within this is international call termination, where the network of a caller in one country must pay the relevant network in the destination country to deliver

[2] Another survey on the same topic can be found in Chapters 3, 4 and 5 of Laffont and Tirole (2000). Points of similarity and contrast between the two surveys will be discussed at various places in this chapter.

its calls (these payments being known as "settlement rates" in this context).

- Call origination: A firm may have chosen not to construct a link directly to its customers, and so must rely on another firm to provide this link. An obvious example is long-distance competition, where a firm may choose to compete only on long-distance services, and therefore needs to pay an incumbent provider of local exchange services to originate (and terminate) its calls.
- Leasing network elements: A firm may lease some part of another firm's network—for instance, the local link into a subscriber's home (known as "local loop unbundling")—for some specified period. At one level this is just like call termination and origination except that the network is leased per month rather than per minute. However, contracts for local loop unbundling might allow the entrant to be more flexible in the technology it uses in combination with the network element—such as a different signalling technology—and so in practice this may be quite different.
- Roaming on mobile networks: For mobile networks, the equivalent of call termination and origination is "roaming", which is when a subscriber moves outside the area covered by her own network A, and still wishes to make and receive calls. In this case network A needs to arrange for a second network with adequate coverage to pick up and deliver the calls.
- Spectrum rights: A more esoteric "access" service is the need by mobile networks for suitable electromagnetic spectrum, a resource which is typically managed by government. The trend is for this essential facility to be auctioned off by government.

2. One way access pricing

Many of the issues involving access pricing are best analyzed in a one way access framework, i.e. where the incumbent firm has a monopoly over important inputs needed by its rivals, but it itself needs nothing from other firms. (Outside the telecommunications sector, in markets such as gas and electricity, the one way paradigm is essentially the only one that is needed.) We discuss one way access pricing policy in a variety of situations. However, the analysis is entirely concerned with the case of vertical integration: the supplier of access services also operates in the retail markets supplied (potentially) by its rivals. The case of vertical separation is simpler, and has been analyzed elsewhere[3]. In addition, the focus is entirely on regulating access *charges*, and the important possibility that the incumbent will try to disadvantage rivals using various non-price instruments is ignored[4].

[3] See for instance, Section 5.2.1 in Armstrong, Cowan, and Vickers (1994) and Section 2.2.5 in Laffont and Tirole (2000). Papers that discuss the disirability or otherwise of vertical integration include Vickers (1995), Weisman (1995), Sibley and Weisman (1998) and Lee and Hamilton (1999). (These latter papers however discuss unregulated downstream markets.)

[4] For further discussion of non-price discrimination against rivals, see Section 4.5 in Laffont and Tirole (2000), Sibley and Weisman (1998), Economides (1998) and Mandy (2000).

Before we present the main analysis, some important preliminaries are discussed: how to achieve efficient entry when the incumbent's retail prices are out of line with its costs (Section 2.1), and how an unregulated vertically integrated firm will behave towards its rivals (Section 2.2). Although it is usual to treat the (regulated) access pricing problem and the (unregulated) foreclosure problem separately, it is one of the pedagogical aims of this chapter to show how many of the principles that underlie the policies that enable an unregulated firm to extract maximum profit from consumers and its rivals carry over to regulatory policies aimed at maximizing social welfare.

2.1. The effect of unbalanced retail tariffs

Incumbent telecommunications firms are often forced to offer retail tariffs that depart significantly from their underlying costs in several dimensions. (As we will see, the access pricing problem would be trivial if this were not the case.) There are two broad ways in which prices can depart from marginal costs. First, there is the problem caused by fixed and common costs: setting all prices equal to marginal cost will not allow the incumbent to break even. This Ramsey problem—which involves the calculation of the *optimal* departures of prices from costs—is discussed in Section 2.5 below. Second, there is the problem that prices are determined in some, perhaps *ad hoc*, manner and may not reflect costs at all, and profits from one market are used to subsidize losses in others[5].

Examples of this practice in telecommunications include:

- A requirement to offer geographically uniform retail tariffs even though the costs of network provision vary across regions;
- A requirement to offer specific groups of subscribers subsidized tariffs;
- A requirement that call charges be used partly to cover non-traffic sensitive network costs, and that fixed charges do not cover the full costs of fixed network provision;
- One-off connection costs are not charged to subscribers as a one-off lump-sum, but much of the cost is collected periodically (e.g. quarterly) as fixed charges;
- Making a call often involves only a call set-up cost and the call cost is independent of duration, and yet call charges are typically levied on a per-minute or per-second basis.

Naturally such patterns of cross-subsidy lead to difficulties with laissez-faire entry, and there will tend to be "too much" entry into artificially profitable segments and "too little" in the loss-making markets. In addition there is the funding problem: if entry eliminates profits in the previously profitable markets,

[5] It is outside the scope of this paper to discuss *why* such cross-subsidies are so prevalent, or whether they are desirable. See Chapter 6 of Laffont and Tirole (2000) and Riordan (2002) for discussions of this, and for further references.

then the incumbent may be unable to continue to fund its operations in loss making markets. Since they have nothing to do with the presence of essential facilities *per se*, for maximum clarity I examine these issues in this section assuming that entrants do not need access to the incumbent's network to provide their services[6].

In the next two sub-sections I discuss how efficient entry can be ensured using two different models for the competitive process. (These two models will be used throughout Section 2 of the chapter.) Despite the apparent duplication, I use both models as they illustrate complementary aspects of competition in the market. The first model might be termed the "unit demands" (or "contestable") model, and it involves a one-for-one displacement of services from the incumbent to the entrant when entry occurs. The crucial aspect of this model is that, as long as prices do not exceed their gross utility, consumers have fixed and inelastic demands. This feature implies that there are no welfare losses caused by high retail prices (provided consumers are served). As a result, we will see that the incumbent will typically not act to distort the market, even if left unregulated, and this feature perhaps has unrealistic welfare implications. In addition, competition is "all or nothing" and either the incumbent *or* the entrant will serve a given subscriber group. This model is used mainly because it provides easy illustrative examples, and provides a simple intuitive way to discuss many of the main policy issues without getting distracted by demand elasticities, product differentiation, and so on.

The second model is termed the "competitive fringe" model, and involves the same service being offered by a group of entrants, where this service is differentiated from the incumbent's own offering. Competition within the fringe means that prices there are driven down to cost and the fringe makes no profit. This model is more appropriate when entry is not all-or-nothing, so that the incumbent retains some market share even when entry occurs. Thus, this model works quite well when discussing competition in the long-distance or international call markets, when a large number of firms may enter the market and yet the incumbent retains a market presence. Because demand curves are downward sloping with this model, there are welfare losses associated with high prices, and this provides a rationale for regulatory intervention[7].

[6] Thus, this section would apply naturally to postal service. See Crew and Kleindorfer (1998) for a related analysis.

[7] Note that neither of these models allow the entrant(s) to have any effective market power, and this assumption is maintained for most of the chapter. (In Section 4.2 on two way interconnection, there are several cases where all firms have market power.) If entrants do have market power then access charges should be chosen with the additional aim of controlling the retail prices of entrants. This would typically lead to access charges being set lower than otherwise, following the same procedure as the familiar Pigouvian output subsidy to control market power. Allowing for this will make an already complex discussion more opaque. However, see Section 3.3.1 of Laffont and Tirole (2000) for a discussion of the implications for policy when entrants have a degree of market power.

2.1.1. Unit demand model

Consider a specific subscriber group that is offered a retail package by the incumbent, denoted M, which may be out of line with M's costs[8]. Suppose M incurs a total cost C per subscriber, and generates gross utility U per subscriber. The price for M's service is mandated to be P per subscriber (this price being determined by a process outside the model). A subscriber's net utility is therefore $U - P$. Suppose there is a potential entrant, denoted E, who can supply a rival package that costs c per subscriber and generates gross utility of u per subscriber[9]. Welfare per subscriber, as measured by the sum of consumer utility and profits, is equal to $u - c$ if E serves subscribers and $U - C$ if M retains the market. Therefore, successful entry is socially desirable if and only if

$$C \geqslant c + [U - u]. \tag{1}$$

Given M's price P, the entrant can attract subscribers provided its own price p is such that $u - p \geqslant U - P$. Therefore, entry will occur whenever the maximum price that can be charged by E, which is $p = P - [U - u]$, covers its costs, i.e. when

$$P \geqslant c + [U - u]. \tag{2}$$

Whenever $P \neq C$, therefore, private and social incentives for entry differ.

There are two kinds of market failure, depending on whether the sector is profitable or loss-making for the incumbent. Suppose first that the market segment is required to be profitable, so that $P > C$. Then whenever

$$P \geqslant c + [U - u] \geqslant C$$

entry occurs when it is socially undesirable. In this case entry can profitably take place even when the entrant has higher costs and/or lower quality than the incumbent. (In this sense there is "too much" entry.) Alternatively, if $P < C$ then when

$$P \leqslant c + [U - u] \leqslant C$$

it is socially desirable for entry to take place, and yet it is not privately profitable. In this case entry does not occur even though the entrant has lower costs and/or a

[8] This discussion, as well as those in Sections 2.3.2 and 2.4.1, is taken from Armstrong (2001). The geographical example is used because it is the simplest way to make the main points. However, the same broad principles apply to other ways of partitioning subscribers, such as into high-usage and low-usage groups. The complication of this kind of alternative framework is that firms will most likely not be able directly to observe the group to which subscribers belong, and so will have to *screen* subscribers—see Section 3.2.3.3 of Laffont and Tirole (2000) for analysis along these lines.

[9] This utility could be higher than that supplied by M (if E uses a newer and superior technology for instance), or it could be lower (if subscribers incur switching costs when they move to the entrant).

higher quality of service. In sum, whenever the incumbent's prices are required to depart from its costs of serving subscribers, there is a danger of undesirable entry into profitable markets and of too little efficient entry into loss-making markets.

In theory it is a straightforward matter to correct this divergence between the private and social incentives for entry. For these subscribers the incumbent is implicitly paying an "output tax" of

$$t = P - C \tag{3}$$

per subscriber—which is positive or negative depending on whether the segment is profitable—and efficient entry is ensured provided that the entrant is *also* required to pay this tax[10]. Notice that this tax is equal to the incumbent's lost profit—or "opportunity cost"—when it loses a subscriber to its rival.

While it may seem a little abstract to use these kinds of output taxes to correct for allocative inefficiencies in the incumbent's tariff, these can sometimes be implemented in a simple and non-discriminatory way via a well designed "social obligations" or "universal service" fund of the kind sometimes proposed for the telecommunications industry. This procedure can be illustrated by means of a simple example, summarized in Table 1 below. (I return to variants of this example later in the chapter.)

Here, the incumbent offers a retail service to two groups of subscribers, a high cost rural group and a low cost urban group. Regulatory policy, in the form of universal service obligations, demands that the incumbent offers service to both groups at the same price, \$100, and (without entry) the firm makes a profit from urban subscribers that just covers the loss from rural subscribers.

As discussed above, a *laissez-faire* entry policy will most likely lead to (i) inefficient entry into the artificially profitable urban market, (ii) too little efficient entry into the loss-making rural sector, and (iii) funding difficulties for the incumbent in the event of cream-skimming urban entry. To counter these problems, suppose the regulator sets up a fund containing \$1bn to finance the rural sector. The source of this fund is the profits generated in the urban sector, and either firm—the entrant or the incumbent—must pay an amount \$50 (M's profit margin in this sector) into this fund for each urban subscriber they serve. In return, any firm that serves a rural subscriber receives a subsidy from the fund equal to \$100 (M's per-subscriber loss in that sector) for each subscriber served. By construction, provided the number of subscribers in the two groups does not change with entry, such a fund is self-financing, and widespread entry does not undermine the ability of the incumbent to supply the loss making market. More important from an economic efficiency point of view, though, is the feature that

[10] With this tax an entrant will find it profitable to take over a subscriber provided that $u - c - t \geqslant U - P$, i.e. whenever $u - c \geqslant U - C$. Therefore, entry takes place if and only if it is desirable.

Table 1
Giving correct entry incentives via a universal service fund

	Urban	Rural
Number of subscribers	20 m	10 m
M's cost per subscriber	\$50	\$200
M's price per subscriber	\$100	\$100
M's overall profit in each sector	\$1bn profit	\$1bn loss
Any firm's contribution to fund	\$50	\$-100

the contribution scheme ensures that in each sector the entrant has to pay the output tax/subsidy (3) which gives it the correct entry incentives[11].

2.1.2. Competitive fringe model

The second model of competition shows how this policy applies in a different context[12]. As discussed above, in order to introduce product differentiation, and to sidestep the issue of the market power of entrants, I assume there is a competitive fringe of price-taking entrants, still denoted E, all of whom offer the same service (which is differentiated from that supplied by M). The impact of modelling competition as a competitive fringe is that we may assume the rivals' price is always equal to their perceived marginal cost, and the fringe makes no profits. Let P and p be M's price and E's price for the respective services. Let $V(P,p)$ be consumer surplus with the two prices. This satisfies the envelope conditions $V_P(P,p) = -X(P,p)$ and $V_p(P,p) = -x(P,p)$, where X and x are, respectively, the demand functions for the incumbent's and the fringe's services[13]. Assume that the two products are substitutes, so that $X_p \equiv x_P \geqslant 0$. As before, the incumbent has marginal cost C and the fringe has marginal cost c. In order to achieve the correct amount of entry, the regulator levies a per unit output tax t on the fringe's service. Then competition implies that the fringe's equilibrium price is $p = c + t$. Keeping the incumbent's price fixed at P, the regulator aims to maximize total surplus (including tax revenue) as given by

$$W = \underbrace{V(P, c+t)}_{\text{consumer surplus}} + \underbrace{tx(P, c+t)}_{\text{tax revenue}} + \underbrace{(P-C)X(P, c+t)}_{M\text{'s profits}}. \tag{4}$$

[11] Clearly, in implementation great care must be taken to ensure that entrants cannot "bypass" this output tax, for instance by providing a similar but not identical service.
[12] This model of competition is adapted from Laffont and Tirole (1994) and Armstrong, Doyle, and Vickers (1996).
[13] Subscripts are used to denote partial derivatives. I assume that consumers have no "income effects", so that Roy's identity reduces to these envelope conditions.

Maximizing this with respect to t implies that the optimal fringe price and output tax is given by

$$p = c + \sigma_d(P - C)\,;\; t = \sigma_d(P - C)\,, \tag{5}$$

where

$$\sigma_d = \frac{X_p}{-x_p} \geqslant 0 \tag{6}$$

is a measure of the substitutability of the two products: it measures how much the demand for M's service decreases when E supplies one further unit of its own service[14]. In particular, if the market is profitable, i.e. when $P > C$, then it is optimal to raise the rivals' price above cost as well, i.e. to set $t > 0$. The reason for this is that profits are socially valuable, and given $P > C$ it is optimal to stimulate demand for M's service, something that is achieved by driving up the rival price (given that services are substitutes).

Analogously to (3) above, t is equal to M's lost profit caused by the *marginal* unit of fringe supply. This lost profit is the product of two terms: the marginal profit $(P - C)$ per unit of final product sales by the incumbent, and σ_d which is the change in the incumbent's sales caused by increasing fringe output by one unit. Therefore, (5) indeed gives the loss in the incumbent's profit caused by E's marginal unit of supply. With perfect substitutes we have one-for-one displacement of rival for incumbent services, i.e. $\sigma_d = 1$, in which case the rule reduces to the simple formula (3). If the products are not close substitutes, so that σ_d is close to zero, then this optimal tax should also be close to zero, and a laissez-faire entry policy is almost ideal. This is because policy towards the fringe market has little impact on the welfare generated in the incumbent's market, and therefore there is no incentive to set a price in the fringe market different from cost.

2.1.3 Discussion: The theory of the second-best

The rules (3) and (5) are instances of the "general theory of the second best", which states that if one service is not offered at the first-best marginal cost price $(P \neq C)$, then the optimal price in a related market also departs from marginal cost $(p \neq c)$[15]. Therefore, taxes of the form (3) or (5) are examples of what might be termed the "second-best output tax". Indeed the optimal policy in this context states that the correct departure from marginal costs should be given by the

[14] Given P, if E supplies a further unit of service its price must decrease by $1/x_p$, which in turn causes M's demand to fall by X_p/x_p. Note that in general σ_d is a function of the tax t, and so expression (5) does not give an explicit formula for the tax. However, provided the demand functions are reasonably close to being linear—so that σ_d is close to constant—this issue is not important.

[15] See Lipsey and Lancaster (1956).

incumbent's lost profit—or opportunity cost—due to entry into the market. We can sum up this discussion by the formula:

$$\textit{Second-best output tax required to implement efficient entry} = \textit{incumbent's lost profit}. \tag{7}$$

From an economic efficiency point of view it makes little difference whether the proceeds from this tax are paid directly to M, to the government, or into an industry fund. However, if the incumbent has historically been using the proceeds from a profitable activity to finance loss-making sectors, then, if the entrant pays the tax to the incumbent, the incumbent will not face funding problems if entry takes place. However, perhaps a more transparent mechanism would be for a "universal service" fund to be used to finance loss-making services, as illustrated by Table 1 above. Notice that if the charge *is* paid to the incumbent firm then the firm is indifferent to whether it retains or loses market share to a rival. This has the political benefit that, with such a tax regime, the incumbent has no incentive to lobby against entry[16]. However, from an efficiency point of view this is irrelevant, and it is merely a happy feature of the regime that this tax which gives correct entry incentives also recompenses the incumbent for its loss of market share.

2.2. The problem of foreclosure in an unregulated market

Does a vertically integrated firm with a monopoly over vital inputs have an incentive to distort competition in related markets? The so-called foreclosure (or essential facilities) doctrine says *Yes*, whereas the so-called Chicago critique claims *No*[17]. It is not the aim of this section to probe this general issue in any depth. Rather, I wish to demonstrate that the main principles that govern the unregulated firm's behavior toward access pricing—or wholesale pricing as it is usually termed in the unregulated case—are closely related to the regulator's incentives when choosing the access charge. In particular, the number of instruments available to the firm/regulator affects whether or not productive efficiency is desirable.

[16] Or at least this is true in the short run; in the longer term the incumbent might believe that entrants will enjoy future benefits (at the incumbent's expense) once they have a presence in the market. In addition, the fact that the incumbent's profits do not depend on the extent of entry implies that it will not have any (short-run) incentive to foreclose entry through non-price means.

[17] For a discussion of the topic of market foreclosure, Rey and Tirole (1996) and Vickers (1996). In this paper I look at the incentive for an already vertically integrated firm to distort the downstream market, whereas most discussions of foreclosure concentrate on the case of vertical separation and the inefficiencies that result from this. (Vertical integration is then often a *solution* to the problem caused by these inefficiencies.)

Consider this simple framework: there is one vertically integrated firm M and a rival firm (or group of firms) E which may need access to M's "upstream" input to be able to compete with M at the "downstream" retail level. Firm M has constant marginal cost C_1 for providing its end-to-end retail product and C_2 for providing its upstream access product to E. Firm M chooses the per-unit charge a for its access service[18]. As in Section 2.1 above, we divide the analysis into two parts, depending on the nature of the downstream competitive interaction. In addition, we consider cases where the entrant can "bypass" M's upstream input—i.e. manage to supply its retail service without the use of this input from M—either by finding an alternative supplier for this service or by providing the service itself.

2.2.1. Unit demand model

As in Section 2.1.1, suppose that M's retail service generates gross utility of U per unit, and that E's service (when it uses M's access service) generates utility u. Suppose that it costs E an additional amount c to transform a unit of M's access service into a unit of its retail service.

No possibilities for bypass: Suppose first that E definitely requires M's access service to provide its retail service. Therefore, as in (1) above, it is socially desirable for E to enter whenever $u - [c + C_2] \geqslant U - C_1$, i.e. when

$$C_1 - C_2 \geqslant c + [U - u] \,. \tag{8}$$

Is it in M's selfish interest to distort the market, and to foreclose the rival even if it is efficient for the entrant to serve the market? In this special framework the answer is *No*. First notice, as in (2), that if M chooses the pair of prices P and a, where the former is its retail price and the latter is the access charge, then the entrant will choose to enter if and only if the margin $P - a$ satisfies

$$P - a \geqslant c + [U - u] \,. \tag{9}$$

(If M chooses the retail price P then to attract subscribers E must offer a price p such that $u - p \geqslant U - P$. It will choose to enter if a price that covers its total cost $c + a$ satisfies this condition.) If M does foreclose E, for instance by setting a very high access charge a, then its maximum profit is obtained by charging the maximum acceptable retail price, which is $P = U$, and this generates profits of $U - C_1$ per unit. On the other hand, if it allows E to provide the retail product, for instance by setting a very high retail price P, then it can set an access charge up to $a = u - c$ (which just allows E to break even at its maximum possible price $p = u$),

[18] More generally, M will want to choose more complicated access tariffs, including franchise fees and so on, to extract any available profits from E. However, in our simple models the rivals necessarily make no profit and so there is no need to use additional instruments for this purpose.

which gives M a profit of $u - c - C_2$ per unit. Therefore, by comparing these profits with the condition for efficiency in (8) we see that M will allow entry if and only if this is efficient. The incumbent manages to obtain the full surplus, and subscribers and E are left with nothing.

Thus the firm has no incentive to distort competition downstream. Essentially this is an instance of the "Chicago critique" of foreclosure, which in simple terms states that vertical integration has no anti-competitive effects. In this case, because of its bargaining power, the unit-demand nature of subscriber demand and the assumption that M has full information about the entrant's technology, M can obtain E's retail service at marginal cost, and so will use this segment whenever its own (quality adjusted) cost is higher[19]. However, as is well known, this result is non-robust in a number of ways—see Section 2.2.2 below for one example. One way in which this insight does extend, though, is to the potential for bypass on the part of the entrant.

Bypass. Suppose next that the entrant can provide its own access service, thus bypassing M's network altogether[20]. Suppose that when it provides its own access it incurs total cost $\hat{C}_1$ per unit for its end-to-end service. This new network gives subscribers a gross utility $\hat{u}$. (Utility $\hat{u}$ may differ from u if using the incumbent's network degrades or enhances the entrant's service compared to its stand-alone service.) Therefore, the entrant can charge a price $p = P + [\hat{u} - U]$ and attract subscribers, without any need for access to M's upstream service. Total surplus per subscriber with this mode of entry is just $\hat{u} - \hat{C}_1$. In sum, social welfare per unit with the three possible entry strategies is

$$W = \begin{cases} \hat{u} - \hat{C}_1 & \text{with stand-alone entry} \\ u - c - C_2 & \text{with entry using } M\text{'s access service} \\ U - C_1 & \text{with no entry.} \end{cases} \tag{10}$$

Given the incumbent's pricing policy (P, a), which of the three options will the entrant follow? If it decides to use M's access service it can charge up to $p = P + [u - U]$ and still attract customers, in which case it makes a profit per unit of $P + [u - U] - [a + c]$. On the other hand, if E bypasses M's network it can make profit of $P + [\hat{u} - U] - \hat{C}_1$. Therefore, *given* that E enters we deduce that it will use M's network provided that

$$a \leqslant [u - \hat{u}] + [\hat{C}_1 - c] \,, \tag{11}$$

[19] Exactly the same efficiency effect is at work when a text-book monopolist faces consumers with unit demands (with known reservation prices). The monopolist will simply charge the maximum price consumers will pay, which causes no deadweight losses due to the shape of the demand functions.

[20] I ignore the possibility that the entrant should build a network and the incumbent should provide retail services over this new network, i.e. that the entrant should provide "access" to the incumbent.

regardless of M's retail price. (The price P affects the profitability of entry itself, but not the relative profitability of the two modes of entry.) It follows that the maximum profit per unit that M can make when it succeeds in supplying access to E is $[u-\hat{u}]+[\hat{C}_1-c]-C_2$.

On the other hand, if it does not supply access to E what is M's maximum profit? Given that E can enter at cost $\hat{C}_1$ without the need for access, the maximum retail price that M can charge and yet not induce stand-alone entry is $P=\hat{C}_1+[U-\hat{u}]$, which gives it profits of $[U-\hat{u}]+[\hat{C}_1-C_1]$. In sum, M's maximum profit per unit under the three entry regimes are

$$\Pi=\begin{cases}0 & \text{with stand-alone entry}\\ [u-\hat{u}]+[\hat{C}_1-c]-C_2 & \text{with entry using } M\text{'s access service}\\ [U-\hat{u}]+[\hat{C}_1-C_1] & \text{with no entry.}\end{cases} \tag{12}$$

Therefore, comparing (12) with (10) we see that the incumbent's incentives are precisely in line with overall welfare. In sum, in this model the potential for bypass does not affect the incumbent's incentive to allow the most efficient method of production. However, bypass *does* affect the profits that M can obtain: comparing (12) with (10) we see that M cannot appropriate all the surplus in (10), and $\hat{u}-\hat{C}_1$ of this surplus leaks out (either to E or to subscribers, depending upon whether entry takes place or not).

In the next section, with the alternative model of competition, we will see that the possibility of bypass *does* cause the incumbent to distort competition, at least when the incumbent can use only a limited set of instruments.

2.2.2. Competitive fringe model

Suppose next that E is a competitive fringe as in Section 2.1.2 above. Suppose that with the access charge a the fringe has the (minimum) constant marginal cost $\psi(a)$ for producing a unit of its final service, including the payment of a per unit of access to M. If the fringe cannot bypass M's access service, so that exactly one unit of access is needed for each unit of its final product, then $\psi(a)=a+c$, say, where c is E's cost of converting the input into its retail product. Otherwise, though, E can substitute away from the access product, in which case $\psi(a)$ is concave in a.

Again we write P for M's retail price and p for E's retail price, and the demand for the two services is $X(P,p)$ and $x(P,p)$ respectively. Note that $\psi'(a)$ is, by Shephard's Lemma, E's demand for access per unit of its retail service, and therefore the total demand for M's access service is $\psi'(a)x$. Suppose for now that M can somehow levy a per-unit charge t on the output of the fringe. Since E is a competitive fringe its equilibrium retail price is equal to total perceived marginal costs: $p\equiv t+\psi(a)$.

Putting all of this together implies that M's total profits, Π, are comprised of three parts:

$$\Pi = \underbrace{(P - C_1)X(P, t + \psi(a))}_{\text{profits from retail}} + \underbrace{(a - C_2)\psi'(a)x(P, t + \psi(a))}_{\text{profits from access}} + \underbrace{tx(P, t + \psi(a))}_{\text{profits from output levy}} . \tag{13}$$

Alternatively, since $p = t + \psi(a)$ we can think of M choosing p rather than t, in which case this profit becomes

$$\Pi = (P - C_1)X(P, p) + \{p - \psi(a) + (a - C_2)\psi'(a)\}x(P, p) . \tag{14}$$

Clearly, whatever choice is made for the two retail prices, profit is maximized by choosing a to maximize the term $\{\cdot\}$ above, which has the first-order condition $[a - C_2]\psi''(a) = 0$. In particular, whenever there are possibilities for the fringe to substitute away from M's access product, so that $\psi'' \neq 0$, then $a = C_2$ at the optimum. Therefore, marginal cost pricing of access is the profit-maximizing strategy for a vertically-integrated firm when it can levy an output tax on its rivals. (The two retail prices P and p are then chosen by M to maximize its profits in (14).) In this framework, then, M has no incentive to distort rival supply in the sense of causing productive inefficiency[21].

An often more realistic case is where M cannot impose an output tax on the fringe, so that $t = 0$ in (13)[22]. In this case the firm chooses P and a to maximize its profits

$$\Pi = \underbrace{\Pi^R(P, a)}_{\text{profits from retail}} + \underbrace{(a - C_2)z(P, a)}_{\text{profits from access}} \tag{15}$$

where $\Pi^R(P, a) \equiv (P - C_1)X(P, \psi(a))$ is M's profit in the retail sector and $z(P, a) \equiv \psi'(a)x(P, \psi(a))$ is the fringe demand for access. (Throughout the chapter, z is used to denote the demand for access.) The solution to this problem has first-order conditions

$$\Pi_P^R + (a - C_2)z_P = 0; \tag{16}$$

$$\Pi_a^R + z + (a - C_2)z_a = 0 .$$

[21] Obviously, however, since both P and p will be set too high from a welfare perspective, the retail market is distorted in the sense of there being allocative inefficiency.

[22] It is unreasonable to suppose that the entrant can be forced to pay the incumbent an output charge if it chooses to supply its own stand-alone service. However, a potentially reasonable contract offered by the incumbent might take the following form: I offer you access at a per unit tied to your acceptance of paying me a per-unit output charge t. For simplicity, though, we assume here that M cannot levy any form of output charge on the fringe, perhaps because of the possible difficulty in observing or verifying the volume of fringe output.

Notice that $a = C_2$ cannot satisfy these conditions, and the firm will choose to set $a > C_2$[23]. Therefore, when the access charge is its sole instrument it is optimal for M to cause there to be a degree of productive inefficiency within the fringe in order to drive up the rival retail price. In contrast to the unit demands model in Section 2.2.1, the vertically integrated firm here *does* have an incentive to distort the production decisions of its rivals.

2.2.3. Discussion: Instruments and objectives

In general the vertically integrated firm has three objectives when trying to maximize its profits: it wishes (i) to ensure there is productive efficiency, i.e. that industry costs are minimized; (ii) to maximize total industry profit given these minimized costs, and (iii) to extract all industry profit for itself. Consider the competitive fringe model of Section 2.2.2. First, note that (iii) is automatically satisfied since competition within the fringe drives profits there to zero. When the incumbent can levy an output tax on the fringe and so has three instruments—P, a and t— we have seen that the remaining two objectives (i) and (ii) can be achieved. This is not surprising: because of product differentiation, objective (ii) requires two instruments, namely the control of the two retail prices; productive efficiency is obtained by the use of a further instrument, the access charge. By contrast, with only two instruments—for instance, if the output levy is not available—then (i) and (ii) cannot both be satisfied, except in the special case where bypass is not possible. (Clearly, when bypass is impossible then (i) is automatically satisfied.) In general, then, a compromise is needed: the access charge is required to control both productive efficiency (which suggests that the charge be close to marginal cost) and the fringe's retail price (a high access charge leads to high retail prices). The optimal access charge, therefore, lies above cost and there is a degree of productive inefficiency.

The same effects will be seen to be at work in the following sections on regulated markets. The difference is that the regulator, not the firm, is controlling the access and other charges, and wishes to pursue two objectives: (i) to ensure there is productive efficiency, as in the unregulated case, and (ii) to maximize total welfare subject to these minimized costs. Again, if the regulator has enough instruments then both of these objectives can be achieved. In other cases, though, the access charge is required to perform too many tasks at once, and productive efficiency will most likely suffer[24].

[23] It is straightforward to show that $z_a < 0$ and, provided $P > C_1$, that $\Pi_a^R > 0$. Therefore, suppose that $a \leqslant C_2$ solves these first-order conditions. For M to make any profits we must have $P > C_1$ and so $\Pi_a^R > 0$. The second of the first-order conditions is therefore a contradiction, and we deduce that $a > C_2$.

[24] See Section 3.3 in Laffont and Tirole (2000) for a more detailed discussion along these lines.

2.3. Fixed retail prices with no bypass: The ECPR

In this section and Section 2.4, we focus on the problem of how best to determine access charges for a *given* choice of the incumbent's retail tariff. (This retail tariff is assumed to be chosen by some regulatory process outside the model.) From an economic efficiency point of view, it is clearly a superior policy to consider the incumbent's retail prices and access charges simultaneously, since that allows for the various tradeoffs between consumer welfare and productive efficiency to be considered correctly—see Section 2.5 for this analysis. However, it seems worthwhile to analyze this case of fixed and perhaps inefficient retail tariffs, since it is often the case that retail tariffs are not set according to strict Ramsey principles (at least as conventionally applied), and various political or historical considerations often have an crucial impact on the choice of the retail tariff.

2.3.1. Different representations of the ECPR

Probably the single most contentious issue in the theory and practice of access pricing is the status of the so-called *efficient component pricing rule*, or ECPR, which states that the access charge should, under certain circumstances, satisfy the formula[25]:

$$\begin{aligned} \textit{access charge} = {} & \textit{cost of providing access} \\ & + \textit{incumbent's lost profit in retail markets} \\ & \quad \textit{caused by providing access}. \end{aligned} \tag{17}$$

What is meant by this becomes clearer if we use the notation introduced in Section 2.2. Using the previous notation, the incumbent's access charge is a, its retail price is P, its marginal cost of supplying its downstream service is C_1, and its marginal cost of providing access to its rivals is C_2. Then the expression (17) can be expressed more formally as

$$a = C_2 + \underbrace{\sigma(P - C_1)}_{\text{M's lost retail profit}} \quad . \tag{18}$$

Here the parameter σ measures how many units of M's retail service are lost by supplying a unit of access to its rivals[26]. In fact we have already seen this rule used implicitly in Section 2.1. In that section there was no vertical element present, and

[25] This rule appears to have been proposed first in the pioneering analysis in Willig (1979)—for instance, see his expression (72). See also Baumol (1983), Baumol and Sidak (1994a), Baumol and Sidak (1994b), Baumol, Ordover, and Willig (1997) and Sidak and Spulber (1997a) for further discussions concerning the desirability of this rule.

[26] The parameter σ was termed the *displacement ratio* by Armstrong, Doyle, and Vickers (1996).

so there was no direct cost of providing access, i.e. $C_2 = 0$. In the unit demands framework, one unit of "market access" enabled one unit of rival service to be supplied, which in turn displaced one-for-one a unit of the incumbent's retail service. Therefore, $\sigma = 1$ in this case and the ECPR proposal reduces to (3). Similarly, in the competitive fringe model, one unit of market access again enabled a unit of fringe service to be supplied, but this caused only σ_d units of incumbent service to be given up, which then gives (5).

From this perspective, using the identity (7) we can re-write (17) to give the perhaps less "incumbent friendly" formula:

$$\begin{aligned} \textit{access charge} = {} & \textit{cost of providing access} \\ & + \textit{second-best output tax on entrants}. \end{aligned} \tag{19}$$

Although it is largely an issue of semantics, this way of expressing the formula better reflects the fact that departures from cost-based access pricing are the result of second-best corrections to account for the incumbent's distorted retail tariff, rather than the result of the need to compensate the incumbent for lost profit. The equivalence between (17) and (19) explains why many of the implications of the analysis in this chapter coincide with those generated from the very different approach of Sidak and Spulber (1997a). I discuss policy from the point of view of static economic efficiency, whereas Sidak and Spulber emphasize more the need to compensate the incumbent adequately for past investments when its network is opened up to rivals[27].

Another formula that often goes under the name of the ECPR, but which is perhaps better termed the "margin rule", states that

$$\begin{aligned} \textit{access charge} = {} & \textit{incumbent's retail price} \\ & - \textit{incumbent's cost in the retail activity}. \end{aligned} \tag{20}$$

Re-arranging, this rule states that the "margin" available to the rivals, i.e. the incumbent's retail price minus its access charge, is equal to the incumbent's cost in the competitive activity. Using the above notation, if C_2 is M's marginal cost of supplying access to itself (as well as to its rivals), then $[C_1 - C_2]$ is M's marginal cost in the competitive activity. Therefore, (20) is more formally expressed as

$$a = C_2 + [P - C_1] \,. \tag{21}$$

Clearly this formula coincides with (18) in the case where one unit of access supplied to the rivals causes a unit reduction in the demand for the incumbent's retail service, i.e. when $\sigma = 1$. Although it is again merely a question of semantics

[27] However, this focus on avoiding "deregulatory takings" could itself be viewed as promoting *dynamic* efficiency—see pages 214–216 in Sidak and Spulber (1997a).

which of the rules (17) or (20) is termed the "ECPR", it is very important when it comes to discussing the applicability of the ECPR to distinguish between these two rules. Since we will see that the former is much more generally valid, we will use the term "ECPR" for that rule, and the term "margin rule" for the latter[28].

We have already seen in Section 2.2 that the level of the optimal access charge is going to depend crucially upon whether the regulator can impose an output tax on entrants. We will see that, suitably interpreted, the ECPR as expressed in (17) or (19) is the correct rule when additional instruments, such as output taxes on entrants, are not employed. When output taxes are used, however, then the ECPR is not appropriate, and pricing access at cost is the better policy (just as in the unregulated case in Section 2.2.).

2.3.2. Unit demands and the margin rule

Consider first the unit demands framework of Section 2.2.1 above. The incumbent has constant marginal cost C_1 for its retail service and constant marginal cost C_2 for its access service. Its retail service generates gross utility U to each of its subscribers, and it is required by regulation to charge the retail price P (this price being determined by a process outside the model). As assumed throughout this section, bypass of the incumbent's network by rivals is not possible, and so (by a suitable choice of units) the entrant needs precisely one unit of access for each unit of its retail service.

There will be entry with the access charge a provided condition (9) holds. Total welfare is higher if the entrant supplies the market if and only if condition (8) holds. Therefore, entry incentives coincide with overall social welfare provided the access charge is determined by the "margin rule" (21). As explained above, this formula is the natural interpretation of the ECPR formula in (17) and (19) in this special context. Note that in this scenario where there is no possibility for the entrant to substitute away from M's access service, there is no productive inefficiency caused by setting the access charge above cost, and hence there is no need for additional instruments such as output taxes to achieve the correct outcome.

It is also worth noting that in this model there is actually no need, from an efficiency point of view, to control the incumbent's access charge (provided the

[28] The discussion in Section 3.2.5 of Laffont and Tirole (2000) is confusing from this perspective in that the margin rule and the ECPR are taken to be the same. For instance, the justification of the margin rule on page 119 is given as (using the notation of this paper): "suppose that an entrant diverts one minute of long-distance phone call from the incumbent. This stealing of business costs the incumbent an amount equal to the markup on his long-distance offering, $P - C_1$, plus the marginal cost of giving access, C_2, to the entrant." This discussion assumes, though, that to steal one unit of business from the incumbent requires the rival to supply exactly one further unit of its own supply, whereas in general with imperfect substitutes the rival will need to supply more or less than the incumbent's lost demand. A final point is that in symmetric situations with product differentiation, the Ramsey optimum involves the margin rule (but never the ECPR in our terminology) being satisfied—see Section 2.5.1 below.

incumbent knows the potential entrant's cost c and service quality u). For if M can choose its own access charge, for fixed P, it will choose to allow entry if and only if the maximum profits from selling access to the entrant are higher than the profits obtained when it sells the retail product itself. But the maximum profits available from selling access are obtained by setting a to be the highest value that satisfies (9), which gives it profits of $P - [c + C_2] - [U - u]$ per unit. Comparing this with the profit obtained by preventing access, which is just $P - C_1$, we see that the incumbent will allow entry to occur if and only if (8) holds, i.e. if entry is more efficient. Note that the level of its regulated retail price does *not* bias the incumbent at all in its dealings with its rival since P affects its profits from selling access and from foreclosing entry in exactly the same way. However, it is important for this result that M knows all relevant information about E's service, something that it unlikely to be the case in practice (since entry has not taken place at the time the incumbent has to choose its access charging policy). The advantage of the rule (21) is that it ensures efficient entry for all kinds of entrant, and does not require that the parameters c and u be known.

Finally, we can illustrate this margin rule in a simple extension of the example used in Table 1[29]. Again, there are two groups of subscribers, rural and urban, and the relevant data is summarized in Table 2 below. Here there are two components to providing a final service: the network element and the retail element. The incumbent is assumed to incur the same retail cost for each subscriber group, but its network cost differs across the two groups. (The incumbent's total end-to-end cost for providing service, denoted C_1 in the preceding analysis, is therefore the sum of these two terms.) Here the entrant cannot build its own network and can compete only in the retail segment. As in Section 2.1.1, the incumbent is forced by policy to charge the same amount to both groups, despite the underlying cost differences.

The margin rule (21) implies that the correct network access charge is \$80 per subscriber, regardless of the type of subscriber[30]. This access charge implies that entry will be profitable only if the entrant has retail costs lower than the incumbent's (or provides a superior service). This policy contrasts with a "cost-based" access charging policy, which would require charging for urban access at \$30 and charging for rural access at \$180, a policy that leads to precisely the same problems as indicated in Section 2.1.1. For instance, with a network access charge of \$30 for urban services, an entrant could have a retail cost of as high as \$70 (as compared to the incumbent's retail cost of \$20) and still find entry profitable.

[29] Baumol (1999) provides a similar argument to that in the following discussion.

[30] The fact that both the retail charge and the retail cost are uniform in this example implies that the margin rule access charge will also be uniform. In fact the access charge, \$80 applied uniformly, is just the geographically averaged network cost in this example, a feature that follows from the assumption that the incumbent is regulated to make zero profits overall.

Table 2
The optimality of the margin rule with no bypass

	Urban	Rural
Number of subscribers	20 m	10 m
M's total cost per subscriber, of which	\$50	\$200
retail cost is	\$20	\$20
network cost is	\$30	\$180
M's retail price for service	\$100	\$100
M's network access charge	\$80	\$80

2.3.3. Competitive fringe model

Next, consider the model described in Section 2.2.2 above, simplified so that there is no possibility of bypass. By making a suitable choice of units, we can assume that the fringe needs exactly one unit of access for each unit of final product it supplies. Therefore, the perceived marginal cost of the fringe is $a + c$, where c is E's cost of converting access into its retail product, and so competition within the fringe implies that E's equilibrium retail price is also $a + c$. As in (4) above, welfare with the access charge a is

$$W = \underbrace{V(P, c + a)}_{\text{consumer surplus}} + \underbrace{(a - C_2)x(P, c + a)}_{M\text{'s profits from access}} + \underbrace{(P - C_1)X(P, c + a)}_{M\text{'s profits from retail}}. \tag{22}$$

As in (5), maximizing this with respect to a gives the following expression for the optimal access charge:

$$a = C_2 + \sigma_d(P - C_1), \tag{23}$$

where σ_d is given in (6) above. (By inspection, the regulator's problems in (4) and (22) are identical, where the "tax" t in the former expression represents the markup $a - C_2$ and the tax revenue in the former plays the same role as the incumbent's profits from access in the latter.) Clearly, (23) is the appropriate version of (17), and so the ECPR is again the correct rule in this framework. Following the discussion in Section 2.3.1 above, however, we see that the margin rule (20)–(21) is *not* valid in this framework.

In the special case where the two retail services are approximately independent in terms of consumer demand, so that $\sigma_d \approx 0$, there is no opportunity cost to the incumbent in providing access. In particular, the ECPR rule (23) states that the access charge should then include no mark-up over the cost of providing access. In the telecommunications context, this situation may be relevant for determining the access charges for the use of the incumbent's network for

such services as mobile and many value-added services, the provision of which *might* be expected not to reduce demand for fixed-link voice telephony services significantly.

2.3.4. Discussion

This section has analyzed the case where the incumbent's retail tariff is fixed and where rivals must have a unit of access in order to provide a unit of their retail service. We showed that the optimal access charge in this situation is given by the sum of the cost of access and a "correction factor". From (7) this correction factor can be interpreted in two ways: it is the incumbent's lost profit in the retail sector caused by suppling access to its rivals as in (17), or it is the second-best output tax to correct for the fact that the incumbent's retail tariff does not reflect its costs as in (19). It is perhaps unfortunate that the former representation is more often used, as it gives entrants the excuse to complain that the ECPR acts to "maintain monopoly profits". (However, this complaint has much greater validity when the incumbent's retail price is not regulated—see Section 2.6 below.) This view of the ECPR is extremely misleading. First, the analysis here is concerned with exogenous retail tariffs, and if there is any "monopoly profit" present then this is the fault of regulation and not of the incumbent. Second, within the framework of this chapter the aim of the formula is *not* to recompense the incumbent for lost profit, but rather to give entrants the correct entry signals. For instance, if this opportunity cost element of the access charge were put into general public funds rather than paid to the incumbent, the efficiency aspects of the ECPR would be undiminished. (However, see footnote 27 above.)

Notice that an alternative way to implement the optimum would be to charge for access at the actual cost C_2 and at the same time to levy a second best tax on the *output* of rivals as proposed in Section 2.1. However, in the present case where exactly one unit of access is needed to produce one unit of output, this output tax might just as well be levied on the input. When this fixed relationship between inputs and outputs ceases to hold, however, this convenient procedure cannot be followed. The next section discusses policy in this often more relevant situation.

2.4. Fixed retail prices and bypass

Here we extend the previous analysis to allow for the possibility that entrants can substitute away from the network service offered by the incumbent, so that the demand for access (per unit of final output by the entrants) is a decreasing function of the access charge. We perform the analysis in two stages: first we examine the situation where the regulator has sufficient instruments to obtain the desired outcome; and secondly we discuss the choice of access charge given that this charge is the sole instrument of policy.

2.4.1. The need to price access at cost with enough instruments

In this section we assume that the regulator has enough instruments to implement the desirable outcome—see the discussion in Section 2.2.3[31]. In particular, since the relationship between the entrant's inputs and outputs is not fixed, we suppose that the regulator can control both the price of access and the entrant's retail price. For instance, suppose that the regulator can levy an output tax on the entrants. (See the next section for an analysis of the case where only the former instrument is available.) We will concentrate on the unit demands framework as the competitive fringe model is so similar.

Unit demand model. Here we follow the bypass model outlined in Section 2.2.1, assuming that the incumbent's retail price P is fixed by regulation. We wish to find a regulatory regime that ensures that the maximum value of welfare in (10) above is achieved. Specifically, suppose that E must pay the tax t per unit of its final output and the charge a per unit of M's network services. Following the earlier argument, given that E enters the market in some way, it will choose to use M's network if (11) holds. On the other hand, given that E enters in some way, (10) implies that welfare is higher when E uses M's network if

$$C_2 \leqslant [u - \hat{u}] + [\hat{C}_1 - c].$$

Therefore, *given* that entry occurs, private and social incentives for using M's network are brought into line by choosing $a = C_2$. Setting the network access charge equal to the cost of providing access gives the entrant the correct "make-or-buy" incentives for its network provision[32].

Turning to the choice for t, following the analysis in Section 2.1.1 the ideal output tax is given by $t = P - C_1$ per unit as in (3) above. With these choices for a and t we see that E's profits per unit with its three options for entry are:

$$E\text{'s profit} = \begin{cases} [\hat{u} - U] + [C_1 - \hat{C}_1] & \text{with stand-alone entry} \\ [u - U] + [C_1 - c - C_2] & \text{with entry via } M\text{'s network} \\ 0 & \text{with no entry.} \end{cases}$$

[31] See also Section 8 of Laffont and Tirole (1994) for related analysis. Section 3.2.4 of Laffont and Tirole (2000) discusses the benefits of levying output taxes on entrants and notes that their use would imply that cost-based access charges are optimal. They regard these kinds of output taxes as being "politically unlikely", however. They go on to suggest that these taxes could be repackaged as a tax on the whole industry to make them seem less discriminatory. This suggestion is illustrated in Table 3 in the current chapter.

[32] Several writers loyal to the ECPR approach have suggested that the ECPR is necessary for productive efficiency—see Baumol, Ordover, and Willig (1997) for instance. When bypass is possible, however, it is usually necessary to price access at cost, rather than at the ECPR level, to ensure productive efficiency at the network level.

Just as was the case with M's profits in (12), comparing these profits with (10) we see that E's incentives are now precisely in line with welfare: the entrant will enter when it is optimal to do so, and will choose to use M's network when that is the more efficient mode of entry. Pricing access at cost means that the entrant has the correct make-or-buy incentives for network construction conditional upon entry, and the output tax (3) means that they have the correct incentives to enter (in any way) given that M's retail tariff is distorted. Other policies will cause various kinds of inefficiencies. For instance, if the entrant is permitted to use the incumbent's network at cost C_2 then it will face the correct make-or-buy incentives conditional on entry, but not the correct incentives to enter. Alternatively, if the ECPR charge (21) were imposed then the rival might build its own infrastructure even if it were more efficient for it to use the incumbent's.

As in Section 2.1.1, the output tax element of this optimal policy can often be implemented by means of a suitably designed universal service fund, as described in Table 3. In this example there is a universal service fund that operates just as in Table 1: any firm providing service to an urban subscriber must contribute \$50 to this fund, and any firm offering service to a rural subscriber can receive \$100 from the fund. In addition to these contributions, the entrant can gain access to the incumbent's network at actual cost (not averaged costs as in Table 2). Notice that if the entrant chooses to enter via the incumbent's network its total payment is \$80 per subscriber, just as in Table 2. However, the advantage of splitting the "ECPR" charge into two parts—a cost-based access charge together with an output tax—is that when network bypass is a possibility it is undesirable to make network access charges deviate from the incumbent's network costs.

Competitive fringe model. We discuss this briefly as it so similar to the unit demands case. Working with the model presented in Section 2.2.2 above, suppose

Table 3
Giving correct entry and make-or-buy incentives

	Urban	Rural
Number of subscribers	20 m	10 m
M's total cost per subscriber, of which	\$50	\$200
retail cost is	\$20	\$20
network cost is	\$30	\$180
M's retail price for service	\$100	\$100
M's profit in each sector	\$1bn profit	\$1bn loss
Any firm's contribution to fund	\$50	−\$100
M's network access charge	\$30	\$180

that the regulator fixes the access charge at a and the per-unit output tax on the fringe at t. Then, similarly to expression (13) above, total welfare in this case is

$$W = \underbrace{V(P, t+\psi(a))}_{\text{consumer surplus}} + \underbrace{(P - C_1)X(P, t+\psi(a))}_{\text{profits from retail}} + \underbrace{(a - C_2)\psi'(a)x(P, t+\psi(a))}_{\text{profits from access}} + \underbrace{tx(P, t+\psi(a))}_{\text{tax revenue from output tax}} \tag{24}$$

or writing $p = t + \psi(a)$ this simplifies to

$$W = V(P,p) + (P - C_1)X(P,p) + \{p - \psi(a) + (a - C_2)\psi'(a)\}x(P,p). \tag{25}$$

Although this differs from the unregulated case in (14) by the addition of the consumer surplus term V, this extra term does not affect the choice of a, which again is chosen to maximize the same term $\{\cdot\}$ and which again leads to marginal cost pricing of access: $a = C_2$. Second, maximizing (25) with respect to $p = t + \psi(a)$ yields the formula (5) for t. Therefore, we see again that, when the regulator can use both these instruments, access should be priced at cost, and the entrants' output tax should be the "second-best output tax".

Thus we have obtained one of the main points in the chapter: just as in the unregulated case discussed in Section 2.2 above, provided there are enough policy instruments available to pursue all the objectives, there is no need to sacrifice productive efficiency even when the incumbent's retail is not cost-reflective. Retail instruments—perhaps in the form of a carefully-designed universal service fund—should be used to combat retail-level distortions such as mandated tariffs that involve cross-subsidies. Wholesale instruments should then be used to combat potential productive inefficiencies—in this case the productive inefficiency caused by pricing access other than at cost[33].

2.4.2. Access charges as the sole instrument: The ECPR revisited

Although one of the main aims of this chapter is to argue that regulators should use output taxes for entrants—perhaps in the guise of a universal service fund—to correct for distortions in the incumbent's retail tariff, and use cost-based access charges to give the correct make-or-buy investment decisions, the former instrument is still only rarely used. Therefore, in this section we consider policy when

[33] This policy suggestion is somewhat related to the "M-ECPR" proposal as outlined in chapter 9 in Sidak and Spulber (1997a). Those authors suggest that the entrant should be charged an amount up to its *own* cost of providing network services for the use of the incumbent's network, and a "competitively neutral end-user charge" should be imposed to prevent cream-skimming entry. (See also Doane, Sibley, and Williams (1999) for further analysis.) One advantage, however, of basing access charges on the *incumbent's* cost is that it decentralizes the decision about the desirability of entry to the (perhaps better informed) entrant, and knowledge of the entrant's technology is not required.

the access charge is the sole instrument available to the regulator[34]. For simplicity, we discuss this issue in the context of the competitive fringe model of Section 2.2.2 above.

In this case we impose $t = 0$ in (24), which yields welfare with the access charge a as

$$W = \underbrace{V(P, \psi(a))}_{\text{consumer surplus}} + \underbrace{(P - C_1)X(P, \psi(a))}_{\text{profits from retail}} + \underbrace{(a - C_2)\psi'(a)x(P, \psi(a))}_{\text{profits from access}}. \tag{26}$$

Maximizing this with respect to a gives

$$a = C_2 + \sigma(P - C_1) \tag{27}$$

where

$$\sigma = \frac{X_p \psi'(a)}{-z_a} \tag{28}$$

and $z(P, a) \equiv \psi'(a)x(P, \psi(a))$ is the fringe demand for access. This formula is again an instance of the ECPR formula (17), suitably interpreted. The first term of the right-hand side in (27) is the direct cost of providing access. The second term is the lost profit to the incumbent in the retail sector caused by providing the marginal unit of access to fringe. This lost profit is itself the product of two terms: M's marginal profit $(P - C_1)$ per unit of final product sales, and the *displacement ratio*, $\sigma = -X_p\psi'(a)/z_a > 0$. The parameter σ gives the reduction in demand for the incumbent's retail service caused by providing the fringe with the marginal unit of access (for a fixed retail price P)[35]. Therefore, the second term in (27) indeed gives the loss in the incumbent's retail profit caused by supplying the marginal unit of access to the fringe.

This formula reduces to no-bypass rule (23) when $z \equiv x$, for in that case $\sigma = \sigma_d$. More generally, it is useful to decompose the displacement ratio σ into two terms, so that

$$\sigma = \sigma_d / \sigma_s,$$

where $\sigma_d = -\frac{X_p}{x_p}$ as in (6) above, and

[34] This section is based on section III of Armstrong, Doyle, and Vickers (1996).

[35] For one more unit of access to be demanded by the fringe, the access charge has to fall by $1/z_a$, and this causes the demand for the incumbent's retail service to fall by $X_p\psi'/z_a$.

$$\sigma_s = \frac{z_a}{x_p \psi'(a)} = \frac{\psi''(a)x + (\psi'(a))^2 x_p}{x_p \psi'(a)} = \psi'(a) + \frac{-\psi''(a)\psi(a)}{\psi'(a)} \frac{1}{\eta_E}, \tag{29}$$

where $\eta_E = -px_p/x > 0$ is the own-price demand elasticity for fringe output. The term σ_d represents the effect of demand-side substitution possibilities, whereas σ_s represents the supply-side substitution possibilities. When there are no supply-side substitution possibilities, so that $\psi(a) = a + c$, then $\sigma_s \equiv 1$ and the displacement ratio is simply $\sigma = \sigma_d$. On the other hand, if the fringe output and the incumbent's output are perfect substitutes then $\sigma_d = 1$ and the displacement ratio is just $\sigma = 1/\sigma_s$.

Expression (29) shows that we may decompose σ_s itself into two terms: the first term $\psi'(a)$ captures the effect of changing a on the demand for M's access service caused by the change in fringe *output* (keeping the input mix constant), while the second term captures the effect of changing a on the demand for M's access service caused by changing the *input mix* (keeping fringe output constant). The term $\psi'(a)$ gives how many units of M's access service is needed for each unit of fringe output. The ability to substitute away from the incumbent's access service is captured by $-\psi''\psi/\psi'\eta_E$ in the above. Since this term is necessarily positive we immediately obtain the basic insight that the ability to substitute away from M's access service causes the displacement ratio to be reduced compared to the no-bypass benchmark (i.e. when $\psi'' \equiv 0$).

Notice that if rival services do substitute for the incumbent's retail service ($\sigma_d \neq 0$) then the ECPR rule in (27) implies that the access charge is not equal to the cost of access, which in turn implies that there is productive inefficiency whenever there is some scope for substitution ($\psi''(a) \neq 0$). The reason for this is that the access charge here is forced to perform two functions, and the regulator must compromise between productive and allocative efficiency. This problem is therefore analogous to that in the unregulated case covered in Section 2.2.2.

2.5. Ramsey pricing

Although Sections 2.3 and 2.4 discussed the important problem of how to use access charges to maximize welfare for a *given* pattern of retail prices imposed on the incumbent, this approach leaves unexamined how these retail prices are chosen in the first place. Therefore, in this section we discuss the problem of optimally choosing the incumbent's retail and access prices simultaneously: the "Ramsey pricing" approach. For simplicity we discuss this problem only in the context of the competitive fringe model. As usual, the form of the solution will depend on whether or not bypass is an option, and, if it is, on the range of policy instruments available to the regulator.

2.5.1. No bypass

Here we extend the model in Section 2.3.3[36]. The problem is to maximize total welfare in (22) subject to the incumbent firm not running at a loss, i.e. that the variable profits—the final two terms in (22)—cover the fixed costs of the firm. Letting $\lambda \geqslant 0$ be the Lagrange multiplier associated with this profit constraint, we see that the retail price P and the access charge a are jointly chosen to maximize the following modification of (22):

$$W = V(P, c+a) + (1+\lambda)\{(a - C_2)x(P, c+a) + (P - C_1)X(P, c+a)\}.$$

Thus, there is now a greater weight placed on the incumbent's profit compared to the previous case, in order to reflect the need for charges to cover fixed costs. Writing $\theta = \lambda/(1+\lambda) \geqslant 0$, the respective first-order conditions for P and a are

$$P = C_1 + \frac{x_P}{-X_P}(a - C_2) + \frac{\theta P}{\eta_M} \tag{30}$$

where $\eta_M = -\frac{PX_P}{X} > 0$ is M's own price demand elasticity, and

$$a = \underbrace{C_2 + \sigma_d(P - C_1)}_{\text{ECPR access charge}} + \underbrace{\frac{\theta p}{\eta_E}}_{\text{Ramsey markup}} \tag{31}$$

where η_E is E's own price demand elasticity and σ_d is as in (6). Since the first two terms on the right-hand side of (31) replicate the corresponding ECPR formula in (23), this formula states that the optimal access charge is the ECPR level, which applies if P were exogenously fixed, plus a Ramsey markup. This Ramsey markup reflects the benefits—in terms of a reduction in P—caused by increasing the revenue generated by selling access to the fringe. In particular, the Ramsey access charge is *above* the ECPR recommendation (which applies taking as given the retail price P). The reason for this is that a higher a raises more revenue that can be used partly to cover the fixed costs, and this allows P to be lowered (which is good for welfare)[37].

It is useful to compare these expressions with the alternative expressions (13) and (18) in Laffont and Tirole (1994), which state (using this chapter's notation) that

[36] This section is based on Laffont and Tirole (1994) and Armstrong, Doyle, and Vickers (1996). See also Section 3.2 of Laffont and Tirole (2000).

[37] Section 8 of Laffont and Tirole (1996) and Section 4.7 of Laffont and Tirole (2000) make the important policy point that Ramsey prices can be implemented by means of a 'global price cap', where both the access and retail services of the incumbent are controlled by means of a suitably designed average price cap. As well as treating wholesale and retail services more symmetrically, they argue that this regulatory mechanism gives the incumbent fewer incentives to exclude rivals by non-price means (compared to, say, cost-based access charges).

$$P = C_1 + \frac{\theta P}{\hat{\eta}_M}, \; a = C_2 + \frac{\theta p}{\hat{\eta}_E}$$

where $\hat{\eta}_M$ and $\hat{\eta}_E$ are the *superelasticities* of the respective products[38]. Although these two pairs of equations look very different, it is possible to show that the expressions (30) and (31) are indeed equivalent to those of Laffont and Tirole. In particular, Laffont and Tirole's expression for the optimal access charge which expresses a as a markup over the cost of providing access involving the super-elasticity, may be re-expressed as we have done in (31), where a is expressed as a markup over the ECPR level involving just the *normal* elasticity—see Section 3.2.2 of Laffont and Tirole (2000) for a similar comparison.

An important issue is the relationship between the Ramsey approach to access pricing and the ECPR approach. Obviously, this question hinges on what we mean by the term "ECPR"—see Section 2.3.1 for various interpretations of this rule. Using our preferred interpretation, which is (23), we see that Ramsey pricing never leads to an ECPR access charge. If, however, one takes the margin rule (21) as the appropriate benchmark then it is indeed possible for the Ramsey access charge to happen to coincide with this margin rule. This issue is discussed in Section 3.2.5 of Laffont and Tirole (2000), who show that with enough symmetry between the incumbent and the fringe the Ramsey access charge does satisfy (21), where P is endogenously determined in the Ramsey problem. However, it is not clear why the margin rule should be a relevant benchmark in this context: with product differentiation the correct opportunity cost incurred by the incumbent in providing a unit of access service is given by (23) and not by (21).

2.5.2. Bypass

Here we extend the Ramsey analysis to allow for bypass by entrants, as in Section 2.2.2. Suppose first that the fringe pays a per-unit output tax equal to t (as well as the per-unit charge a for access input). As usual, the price of the fringe product is equal to the perceived marginal cost, so that $p = t + \psi(a)$, and fringe profits are zero.

As in 2.2.2 and 2.4.1 above, the regulator can be considered to choose p directly rather than t, in which case the incumbent's profit Π is (14). Let $\lambda \geqslant 0$ be the shadow price on the profit constraint $\Pi \geqslant F$, where F is the fixed cost that needs to be covered by profits. Then the problem is to choose P, p and a to maximize $W = V(P, p) + (1 + \lambda)\Pi$. However, since for given retail prices P and p the access charge a does not affect consumer surplus, it is clear that a must be chosen to maximize Π for given retail prices, so that a again maximizes the term $\{\cdot\}$ in (14). As before, provided there are some possibilities for substituting away from the

[38] The reason that the two pairs of first-order conditions look different is that this chapter's formulae are obtained by maximizing over prices, whereas Laffont and Tirole maximize over quantities.

incumbent's access product, the first-order condition for this cost minimization is $a = C_2$ and so pricing access at cost is optimal[39]. This is just an instance of the general result that productive efficiency is desirable when there are enough tax instruments—see Diamond and Mirrlees (1971)[40].

Next, as in Section 2.4.2, suppose the regulator has a more limited set of policy instruments, and that the output tax t is not available. In this case $p \equiv \psi(a)$ and the access charge must perform two functions: it must attempt to maintain productive efficiency (as before) but in addition it must influence the fringe retail price in a desirable way. The incumbent's profit is now as in (15). Writing $\theta = \lambda/(1+\lambda) \geqslant 0$, the respective first-order conditions for maximizing $V + (1+\lambda)\Pi$ for P and a are respectively

$$P = C_1 + (a - C_2)\frac{z_P}{-X_P} + \frac{\theta P}{\eta_I}$$

and

$$a = \underbrace{C_2 + \sigma(P - C_1)}_{\text{ECPR access charge}} + \underbrace{\frac{\theta a}{\eta_z}}_{\text{Ramsey markup}}, \tag{32}$$

where σ is as given in (28), and $\eta_z = -\frac{az_a}{z}$ is the own price elasticity of the demand for access. In particular, the Ramsey access charge is again above the associated ECPR recommendation, which this time is given by (27). These first-order conditions imply that $P > C_1$ and $a > C_2$, and so access is priced above marginal cost. This in turn leads to a degree of productive inefficiency. Just as in Sections 2.2.3 and 2.4.2 the access charge is called upon to perform too many tasks, and a compromise must be made. In the next section the access charge is forced to perform yet another task, which is to try to control the *incumbent's* unregulated retail price.

[39] This paragraph provides one argument, to do with distortions at the input level, for the use of both retail and wholesale instruments for policy. Another rationale for this might be the following: Ramsey principles imply that different retail services that use the same access service should typically have different retail prices, depending on the demand elasticities. When there is no scope for bypass, in some cases one could implement this outcome by differential, use-dependent access charges. In others, though, this may not be possible, perhaps because the incumbent cannot accurately monitor the use to which its network is put. In such circumstances differential retail taxes could be used to implement the Ramsey solution, and non-discriminatory cost-based charges could then be used for network access.

[40] On pages 370–71 of Sidak and Spulber (1997a) those authors try to argue that pricing access differently from cost—as with ECPR-style access charges—does not violate the Diamond-Mirrlees result.

2.6. Unregulated retail prices

In this section we discuss how best to price access when this is the only instrument for regulatory control of the incumbent[41]. In particular, the incumbent is now assumed to be free to set its retail price. For simplicity we suppose that there is no output tax on the fringe. (It would seem strange to consider a situation where the entrants were regulated in some sense while the incumbent was not.)

2.6.1. Perfect retail competition

To discuss this topic in its simplest and starkest form, we first model downstream competition as being perfect. Specifically, there is a group of rivals which offers a service which is a perfect substitute for the incumbent's. Consumer demand for this homogeneous service is denoted $Q(P)$, where P is the price offered by firms in the market. For simplicity suppose the rivals cannot bypass the incumbent's access service, and it costs rivals c to convert a unit of access into a unit of retail product[42]. As usual, the incumbent has marginal cost C_1 for supplying its end-to-end retail service and cost C_2 for providing access to the fringe. Therefore, it is efficient for the fringe to supply the market whenever

$$c + C_2 \leqslant C_1. \tag{33}$$

Control of the access charge. Suppose regulation fixes the access charge at a. Then there will be entry if M chooses its retail price above $a + c$. In this case M will obtain profit $a - C_2$ per unit from providing access to the rivals. On the other hand, if M wishes to stay in the retail market the maximum retail price it can charge is the limit price $P = a + c$, which generates profit $a + c - C_1$ per unit. Therefore, the incumbent will choose to allow entry if and only if this is efficient as in (33). (The level of the regulated access charge does not affect the relative profitability of the two strategies, although it will affect the absolute profitability of either.) We deduce that, when the incumbent is free to set its retail price, it will allow entry if and only if the entrant has lower costs, regardless of the level of the access charge. In particular, because of opportunity cost considerations, the fact that a differs from C_2 does not distort at all the incumbent's incentives to compete on a "level playing field" with its downstream rivals.

Since we have seen that productive efficiency is automatically ensured in this perfect competition setting, the access charge should be chosen to attain allocative

[41] This is adapted from Section 7 of Laffont and Tirole (1994), Section 5.2.2 of Armstrong, Cowan, and Vickers (1994) and Armstrong and Vickers (1998). For other analyses of access pricing with an unregulated downstream sector, see Economides and White (1995), Vickers (1995), Lapuerta and Tye (1999) and Lewis and Sappington (1999). In addition, Sections 3 and 4 of this chapter consider other situations where firms are unregulated at the retail level.

[42] Little would change if we allowed for the possibility of bypass, since this would just strengthen the motive to choose an access charge close to cost.

efficiency. Notice that, regardless of whether entry is successful, the equilibrium retail price given a is $P = a + c$ (at least if this is below the unregulated monopoly retail price). Therefore, we wish to choose a so that this price equals (minimized) marginal cost, so that $a + c = \min\{C_1, C_2 + c\}$. In the case where entry is more efficient this implies that $a = C_2$ and access is priced at marginal cost. In the case where the incumbent should serve the market, we should ideally set $a = C_1 - c$[43]. However, when the market is fairly symmetric, in the sense that $C_1 \approx C_2 + c$, pricing access at marginal cost will be a good approximation to the ideal access pricing regime[44].

Control of the margin. Next, consider the effects of controlling the margin $P - a$ rather than the access charge a, so that M can choose any pair of prices P and a that satisfy $P - a = m$, say. Clearly, entry takes place if and only if this regulated margin covers the fringe's retail cost, i.e. if $m \geqslant c$. In particular, with this regime M has no discretion about whether or not entry takes place. If entry does take place then, given the access charge a, the fringe retail price will be $p = a + c$, and M's profit will be $(a - C_2)Q(a + c)$. If we think of M as choosing the fringe retail price $p = a + c$ rather than a, then p will be chosen to maximize $(p - c - C_2)Q(p)$. On the other hand, if $m < c$ then M will choose its retail price P to maximize its profits $(P - C_1)Q(P)$. In either case we have the outcome corresponding to unregulated monopoly—with entry the marginal cost is $c + C_2$ and with the incumbent serving the retail market the marginal cost is C_1—and this kind of regulation exerts no downward pressure on retail prices whatsoever[45]. However, since welfare with unregulated monopoly increases when the marginal cost is reduced, given that margin regulation is being used, it is better to have entry if and only if it is efficient. Therefore, we want m to be chosen so that $c \leqslant m$ if and only if $c + C_2 \leqslant C_1$. In other words, optimal margin regulation entails $m = C_1 - C_2$, i.e. that the margin rule (21) should be applied.

However, it is simple to verify that this is exactly the outcome when M is totally unregulated. (In Section 2.2.1 we saw in a similar model that the unregulated incumbent will always allow entry when this is efficient, and will then maximize profits.) Thus we see that the *best* form of margin regulation—which is given by the margin rule version of the ECPR—simply replicates the totally unregulated outcome. This provides a compelling argument against the use of the ECPR in deregulated markets. Moreover, this argument gives validity to the common

[43] By construction, this access charge is below the cost of providing access, C_2, in order to eliminate the mark-up that the more efficient incumbent would otherwise be able to charge in this market.

[44] This is essentially a formalization of the argument in Section 7 of Lapuerta and Tye (1999), who argue that access should be priced at cost because competition in the retail sector can be relied upon to eliminate all distortions in that segment.

[45] This has a similar flavor to the duopoly analysis in pages 26–27 of Laffont, Rey, and Tirole (1998a). Those authors find that a form of margin regulation—where networks choose their access charge and then the allowed retail prices must generate a specified margin—facilitates collusion, i.e. that there is again no downward pressure on retail prices.

complaint that the ECPR acts to "maintain monopoly profits"[46]. (But see the discussion in Section 2.3.4 for why this complaint is not valid when the incumbent's retail tariff is controlled by regulation.) This policy contrasts with the preceding case where the access charge is the instrument of policy, and where a low access charge translates directly into low retail prices. Therefore, we can deduce that direct control of the access charge is superior in terms of welfare compared to margin regulation. Indeed, in this perfect competition framework, pricing access at cost—or a little lower in some cases, if feasible—is then the optimal policy.

2.6.2. Competitive fringe model

To investigate this topic in more detail, we return to the competitive fringe model with bypass as in Section 2.2.2. Therefore, the fringe equilibrium retail price is $p = \psi(a)$, and M's profit as a function of P and the access charge a is as given in (15) above. For a given access charge a, the incumbent chooses P to maximize its profits, the solution to which has the first-order condition (16). Note that this may be rearranged to give

$$a = C_2 + \underbrace{\frac{-\Pi_P^R}{z_P}}_{M\text{'s lost retail profit}}. \tag{34}$$

This can again can be interpreted as an instance of the ECPR rule (17) above[47]. The first term on the right-hand side of the above is the cost of providing access, while the second is M's loss in profits in the retail sector caused by supplying the marginal unit of access to the fringe[48]. In addition, since $z_P > 0$ (34) implies that Π_P^R has the opposite sign to $(a - C_2)$. This has a natural intuition: if access is priced above cost, the incumbent has an incentive to push its retail price *above* the profit-maximizing level for the retail sector viewed in isolation since it also wishes to stimulate demand for its profitable access service. (The reverse argument holds if $a < C_2$.)

The profit-maximizing choice of P given a, denoted $\bar{P}(a)$, maximizes Π in (15). In most reasonable case it makes sense to assume that $\bar{P}'(a) > 0$, so that a higher regulated access charge leads to a higher equilibrium retail price for the incumbent: the more profitable selling access to its rivals is, the less aggressively the

[46] This point is forcefully made in section II(A) of Tye and Lapuerta (1996).

[47] One would not want to push this interpretation too far. In Sections 2.3 and 2.4 we made the normative point that the optimal access charge given a specific and fixed retail price was the ECPR. In this section we merely make the *positive* point that, for any given access charge, the unregulated incumbent's choice of retail price happens to have an ECPR flavor.

[48] To cause a further unit of access to be demanded requires, for a given a, that P rises by an amount $1/z_P$, which in turn causes profits in the retail sector to fall by the amount given in the formula.

incumbent will compete with rivals. (The following analysis can easily be adapted if this assumption is invalid.) The welfare-maximizing choice for a is derived as follows. Let $\bar{\Pi}(a)$ be the incumbent's maximum profit in (15) given a. Choosing a to maximize welfare $V(\bar{P}(a), \psi(a)) + \bar{\Pi}(a)$ implies that

$$-X\bar{P}' - x\psi' + \bar{\Pi}' = 0,$$

which, by using the envelope theorem for $\bar{\Pi}'$ and the fact that $z = x\psi'$, implies that

$$a = \underbrace{C_2 + \sigma(\bar{P}(a) - C_1)}_{\text{ECPR access charge}} - \underbrace{\frac{X\bar{P}'}{-z_a}}_{\text{mark-down to control } P} \tag{35}$$

where σ is given in (28). Therefore the access charge should be *below* the ECPR level in (27) given by the situation where M's retail price was *fixed* at $\bar{P}(a)$: the fact that M's retail price is unregulated implies that the access charge should be set below the ECPR level that should apply if its retail price were fixed. The intuition for this is clear. If controlling M's retail price were not an issue, Section 2.4.2 argued that the optimal access charge was given by the ECPR in (27). But now a reduction in a causes the retail price P to fall, which is good for welfare. In sum, because M's retail price is positively related to its access charge, there is a need to reduce the access charge to below the ECPR level[49]. (By contrast, in the Ramsey problem it was optimal to raise the access charge from the ECPR level—see expression (31) above—for the reason that an *increase* in the access charge there caused the retail price to fall, since the access service then financed a larger share of the fixed costs of the firm.)

Another natural comparison is between a and the cost of access C_2. (The fact that a is below the ECPR level says little about the comparison of a with C_2.) In Section 2.6.1, we saw that pricing access at cost—or sometimes a little below cost if this were feasible—was the optimal policy. However, in this more general framework it seems hard to obtain clear-cut results about whether a should be above or below cost, and in general either can be optimal. However, in a few special cases the optimal access charge should precisely equal cost:

- With no possibility for bypass and if the demand functions X and x are linear then the regulator should set $a = C_2$[50].

[49] A similar point is made in section III of Economides and White (1995). They show that when the downstream market is unregulated it can be desirable to allow entry by an inefficient firm—something that is achieved by choosing an access charge below the ECPR level—if this causes retail prices to fall, i.e. it can be a good thing to sacrifice a little productive efficiency to reduce allocative inefficiency.

[50] See Section 7 of Laffont and Tirole (1994) and Armstrong and Vickers (1998).

- If there are no cross-market effects then the regulator should set $a = C_2$. (From (15) the profit-maximizing retail price $\bar{P}$ does not depend on a in this case, and also $\sigma = 0$, therefore (35) implies that marginal cost pricing is optimal.)

However, in general it will simply be by chance that the optimal access charge is equal to cost. As a result we expect that when bypass is a possibility the result will be productive inefficiency.

The reason that it is hard to get clear-cut results in this framework is because the access charge here is called upon to perform *three* tasks: (i) it is used to control the market power of the incumbent (a lower value of a feeds through into a lower value for P); (ii) it is used to achieve allocative efficiency *given* P using the second best argument of Section 2.3.3, and (iii) it is used to try to achieve productive efficiency (which requires $a = C_2$) whenever there is a possibility for bypass. In general, motives (i) and (iii) argue for an access charge no higher than cost. (When $a = C_2$ the incumbent will choose $P > C_1$, and so choosing $a < C_2$ will bring M's retail price down towards cost. Motive (iii) will merely temper but not overturn this incentive.) However, unless a is chosen to be so low that $P < C_1$, motive (ii) will give the regulator an incentive to raise a above cost—see the ECPR expression (27). Therefore, because of these forces pulling in different directions, it is not possible to give clear guidance about the relationship in unregulated retail markets between the access charge and the cost of providing access.

2.6.3. Partial deregulation

In practice the incumbent firm often operates in retail markets that are partially regulated[51]. Thus the firm may have discretion choosing its retail price in a given market, but only subject to certain constraints on the overall pattern of its retail prices. There are two main kinds of constraint that are commonly imposed. First, the incumbent's retail tariff could be controlled by some kind of average price cap, so that it has freedom to vary relative retail prices subject to an overall cap. In this case a decrease in one retail price—say, in response to entry—then allows the firm to increase other retail prices. Second, the firm may face constraints on "price discrimination", so that it is free to determine the *level* of its retail prices but faces constraints on the structure of relative prices (such as a geographically uniform tariff restriction). Here, if the firm decreases one price in response to entry then it is forced to *decrease* other prices in related markets.

How is our analysis modified by these cross-market interactions? Suppose for simplicity there are two retail markets, in one of which the incumbent faces entry and in the other no threat from entry. Suppose also that consumer demand in the two markets is independent. Write P_1 to be M's price in the potentially more

[51] See Section 5 in Vickers (1997) and Section 4.7 in Laffont and Tirole (1999) for some discussion of this topic. This topic is also closely related to the analysis of Armstrong and Vickers (1993), although there are no vertical issues in that paper.

competitive market, and let P_2 be the price in its captive market. As usual, let a be the network access charge. In this case M's profit in (15) is modified to be

$$\Pi = \underbrace{\Pi_2^R(P_2)}_{\text{captive market}} + \underbrace{\Pi_1^R(P_1,a)}_{\text{competitive market}} + \underbrace{(a-C_2)z(P_1,a)}_{\text{access market}} \tag{36}$$

where $\Pi_1^R(P_1,a) \equiv (P_1 - C_1)X_1(P_1,\psi(a))$ is M's profit in the competitive retail market, $\Pi_2^R(P_2)$ is its profit function in the captive market, and $z(P_1,a) \equiv \psi'(a)x(P_1,\psi(a))$ is the fringe demand for the access needed for its activity in the competitive market.

Suppose that M is constrained by policy to choose $P_2 = \zeta(P_1)$. When the firm faces an average price cap, the relationship ζ is decreasing: a rise in P_1 requires a fall in P_2. By contrast, a ban on price discrimination across the two markets could be represented by the constraint $P_2 = P_1$, so that ζ is increasing. If we impose the constraint $P_2 = \zeta(P_1)$ in (36), write $\bar{P}_1(a)$ to be the resulting profit-maximizing price P_1 when the access charge is a. In most reasonable cases we continue to have $\bar{P}_1(a)$ being an increasing function of the access charge: the more profitable selling access to its rivals is, the less aggressively the incumbent will compete when faced with entry. Then (35) is modified to be

$$a = \underbrace{C_2 + \sigma(\bar{P}_1(a) - C_1)}_{\text{ECPR access charge}} - \underbrace{\frac{(X_1 + X_2\zeta')\bar{P}_1'}{-z_a}}_{\text{correction to control } P_1,P_2}. \tag{37}$$

(Here $\zeta' = \zeta'(\bar{P}_1(a))$ and X_1 and X_2 are the incumbent's equilibrium demands in its two retail markets.)

The effect of the cross-market price constraints are then as follows. First there is an effect on the firm's incentive to compete, as captured by $\bar{P}_1$ and $\bar{P}_1'$. In the case of a ban on price discrimination we expect that, because the firm is reluctant to lose profits in its captive market, it is less responsive to changes in the access charge. Indeed, in the limit as profits in the captive market become the incumbent's sole objective then $\bar{P}_1' = 0$ and (37) implies that the usual ECPR formula is then optimal. In effect, the retail price in the competitive market is now fixed exogenously, and the earlier analysis of access pricing with fixed retail prices in Section 2.3 goes through as before. In less extreme cases the presence of the captive market will just temper the incumbent's incentive to vary P_1 in response to policy towards the access charge, and the correction factor on the right-hand side of (37) will correspondingly be reduced.

Secondly, when the firm operates under an average price cap, then typically this correction factor will again be reduced, and the ECPR formula will again be closer to being optimal than indicated in Section 2.3 For instance, suppose that the price cap takes the "fixed weights" form: the incumbent must choose retail

prices satisfying $w_1P_1 + w_2P_2 \leqslant P^*$ for some positive constants w_1 and w_2. In this case $\zeta' \equiv -w_1/w_2$. Suppose further that these fixed weights are chosen to be proportional to the equilibrium demands in the two markets, so that $w_1/w_2 = X_1/X_2$[52]. In this case the correction factor in (37) vanishes entirely, and the ECPR is again valid. In this special case the captive market provides exactly the appropriate correction factor to control the incumbent's retail price in the competitive market, and the access charge can be chosen as if the retail price in the more competitive market was fixed.

2.7. Introducing dynamic issues

Until this point (as indeed it will be for the remainder of the chapter) the analysis has been purely static. To see how this static analysis generalizes easily to encompass some dynamic aspects, consider this simple extension to the no bypass model of Section 2.3.3[53]. Investment, production and consumption take place at discrete points in time, $t = 0, 1, \ldots$, and suppose that the prices for M's product and E's product in period t are, respectively, P_t and p_t. Suppose that consumer surplus, and demand functions for the incumbent's and fringe's products in period t are, respectively, $V_t(P_t, p_t)$, $X_t(P_t, p_t)$ and $x_t(P_t, p_t)$. (We assume there are no intertemporal linkages in demand.) Over time the incumbent invests in a network which supplies access, and suppose that one unit of access is needed to provide one unit of retail service, either by M or by E. Suppose that the capacity of the network at time t, which determines the total number of units of retail service that can be generated by the network, is K_t. Capacity depreciates at the proportional rate δ. There are constant returns to scale in installing capacity, and suppose that installing a unit of capacity in period t costs β_t. Let I_t be the amount of investment (in money terms) in period t, so that the amount of new capacity installed in period t is I_t/β_t. Then capacity evolves according to the dynamic relation

$$K_{t+1} = (1 - \delta)K_t + \frac{I_{t+1}}{\beta_{t+1}}. \tag{38}$$

(We suppose that the current period's investment can be used with immediate effect.)

[52] Section 2.2.2 of Laffont and Tirole (1999) discusses this form of average price regulation in more detail. In particular, they show that, in the absence of competition, this form of price regulation leads to Ramsey-like prices.

[53] The following extends the analysis in Section 4.4.1.3 of Laffont and Tirole (2000). See Hausman and Sidak (1999) and Jorde, Sidak, and Teece (2000) for analyses of the important issue of how dynamic access price regulation affects the incumbent's incentive to innovate and upgrade its network. For a treatment of dynamic issues in access pricing which emphasizes the problem of commitment and expropriation in regulatory policy, see Sidak and Spulber (1996), Sidak and Spulber (1997a) and Sidak and Spulber (1997b).

What is the marginal cost of providing an extra unit of access in period t? Suppose the investment plan is $K_t, K_{t+1}, \ldots, I_t, I_{t+1}, \ldots$ satisfying (38). If K_t is increased by 1, then all subsequent values for K and I are unchanged provided that next period's investment I_{t+1} is reduced so as to the keep the right-hand side of (38) constant, i.e. if I_{t+1} is reduced by $(1-\delta)\beta_{t+1}$[54]. If the interest rate is r, so that \$1 next period is worth $\$\frac{1}{1+r}$ now, then the net cost of this modification to the investment plan is

$$LRIC_t = \beta_t - \frac{1-\delta}{1+r}\beta_{t+1}.$$

(The notation $LRIC$ is used to stand for 'long-run incremental cost', the term often used in policy debates.) If technical progress causes the unit cost of new capacity to fall at the exogenous rate γ every period, then $\beta_{t+1} = (1-\gamma)\beta_t$. With technical progress γ, the above formula reduces to

$$LRIC_t = \beta_t\left(1 - \frac{(1-\delta)(1-\gamma)}{1+r}\right). \tag{39}$$

Notice that if the parameters δ, r and γ are each reasonably small, this formula is approximated by $LRIC_t \approx \beta_t(r+\gamma+\delta)$[55].

Suppose that it costs M and E an amount C_t and c_t respectively to convert a unit of access into their final retail services. Let a_t be the access charge paid by E in period t. Then competition implies that the fringe's price in period t is $p_t = a_t + c_t$, and total discounted welfare is

$$W = \sum_t \frac{1}{(1+r)^t}\{V_t + X_t(P_t - C_t) + x_t a_t - I_t\},$$

where $K_t \equiv X_t + x_t$ and this capital stock evolves according to (38). (All functions in the above are evaluated at the prices P_t and $p_t = a_t + c_t$.) Since (38) implies that

$$I_t = \beta_t[X_t + x_t - (1-\delta)(X_{t-1} + x_{t-1})],$$

substituting this value for investment into the expression for welfare W yields

$$W = \sum_t \frac{1}{(1+r)^t}\{V_t + X_t(P_t - C_t) + x_t a_t - \beta_t[X_t + x_t - (1-\delta)(X_{t-1} + x_{t-1})]\}.$$

[54] I assume that demand conditions are such that investment in each period is strictly positive, which ensures that this modification is feasible.

[55] This is a familar equation in continuous time investment models—see for instance expression (7) in Biglaiser and Riordan (2000).

Maximizing this with respect to P_t and a_t yields

$$\left\{\frac{\partial X_t}{\partial P_t}(P_t - C_t) + \frac{\partial x_t}{\partial P_t}a_t - \beta_t\frac{\partial[X_t + x_t]}{\partial P_t}\right\} + \frac{1}{1+r}\left\{(1-\delta)\beta_{t+1}\frac{\partial[X_t + x_t]}{\partial P_t}\right\} = 0$$
$$\left\{\frac{\partial X_t}{\partial p_t}(P_t - C_t) + \frac{\partial x_t}{\partial p_t}a_t - \beta_t\frac{\partial[X_t + x_t]}{\partial p_t}\right\} + \frac{1}{1+r}\left\{(1-\delta)\beta_{t+1}\frac{\partial[X_t + x_t]}{\partial p_t}\right\} = 0. \tag{40}$$

Together these imply that the access charge should be given by $a_t = LRIC_t$ in (39), and that $P_t = C_t + a_t$. Thus, as expected in this constant-returns-to-scale world with all prices chosen to maximize welfare, it is optimal to set all charges equal to the relevant marginal costs. In particularly, it is optimal at each point in time to set a_t equal to the correctly calculated marginal cost (which falls at the rate of technical progress γ each period).

In order to bring out the similarities with Section 2.3, we next calculate optimal access charges $\{a_t\}$ for a *given* time-path for the incumbent's retail prices $\{P_t\}$. These are derived from expression (40) above, which can be simplified. As in (6), if we write

$$\sigma_t = \frac{\partial X_t/\partial p_t}{-\partial x_t/\partial p_t}$$

for the demand substitution parameter in period t, then (40) simplifies to

$$a_t = \underbrace{LRIC_t}_{\text{access cost in } t} + \underbrace{\sigma_t[P_t - C_t - LRIC_t]}_{\text{lost retail profit in } t}$$

where $LRIC_t$ is given in (39). This is just a dynamic version of the ECPR rule (23) derived above. Thus we see that the formulae generated in the simple static models extend in a straightforward way to simple dynamic models[56].

2.8. Controversies and conclusions

It is hard to find a more controversial issue in industrial policy than that concerning the terms on which entrants can gain access to an incumbent firm's network. In this section I try to summarize the main insights generated by the analysis, and how these can shed light on policy debates.

[56] For further analysis of the dynamic access pricing problem, focussing on the issue of how to encourage multiple networks to provide infrastructure investments at the efficient level, see Gans and Williams (1999) and Gans (2001).

Three broad kinds of access charging policy are:

1. Pricing access at cost,
2. Ramsey pricing, i.e. choosing the incumbent's retail prices and access charges simultaneously to maximize welfare, and
3. the ECPR, i.e. pricing access at the cost of access plus the incumbent's opportunity cost.

Economic opinion on the applicability of these policies often seems to be extremely polarized, and cost-based access charges and ECPR-based access charges both have their fervent supporters. As usual, though, the relative merits of these policies depends on the specifics of each case, and these are discussed in turn.

2.8.1. Cost-based access charges

The chief benefits of cost-based access charges are twofold. First, they are relatively simple to implement. (Or at least as simple as estimating the incumbent's network costs, something which is needed for all reasonable access pricing policies.) In particular, to calculate these charges no information is needed about subscriber demand, nor about the characteristics of entrants (at least in the simple models presented above). Second, this is the only access pricing policy that gives the correct "make-or-buy" signals to entrants when bypass is a possibility. For instance, pricing access above cost could mean that an entrant would prefer to bypass the incumbent's network and construct its own network, even though it would be more efficient to use the incumbent's network. A third, and less clear-cut, benefit of such charges is that they are "fair and non-discriminatory", and do not depend upon the use which is made of the incumbent's network by rivals. Therefore, under this regime different entrants will not be offered different wholesale terms by the incumbent.

In broad terms, cost-based access charges are appropriate when access charges do not need to perform the role of correcting for distortions in the incumbent's retail tariff. There are three main reasons why such a task might not be necessary:

1. First, if the incumbent's regulated retail tariff *does* reflect its underlying costs accurately then no "second-best" corrective measures are needed at all, and in such cases access charges should also reflect the relevant costs. In sum, a full and effective rebalancing of the incumbent's tariff greatly simplifies the regulatory task, and allows access charges to perform the focussed task of ensuring productive efficiency.
2. Second, if there are distortions present in the regulated tariff, but the second-best corrections are made via another regulatory instrument (such as an output tax levied on entrants), then access charges need not perform this additional task, and so again can safely reflect costs—see Sections 2.4.1 and 2.5.2 above. Indeed, one of the main aims of this analysis has been to argue that cost-based

access pricing *is* the best policy, provided that the incumbent's retail distortions are more directly tackled by, for instance, a well-designed universal service fund.

3. Third, when the incumbent operates in a vigorously competitive retail market and is free to set its own retail tariff, we argued that pricing access at cost was a good policy—see Section 2.6.1 above. The fact that the downstream sector is competitive implies that the incumbent has no significant opportunity costs, and so once again access charges should reflect costs.

However, in other cases—i.e. when opportunity cost considerations apply and when access charges are required to correct for these—we have seen that pricing access at cost is sub-optimal.

2.8.2. Ramsey pricing

Almost by definition, Ramsey pricing is the best way to set access (and other) prices. Once the regulator has chosen a measure of social welfare—which could include special care being taken over the welfare of certain subscriber groups—then the optimal policy is to choose access and retail charges to maximize this welfare function, subject to constraints on the profitability of the firm and/or the costs of public funds. This is clearly superior to a policy whereby retail prices are (somehow) chosen first, perhaps without regard for how access charges will subsequently be chosen, and access charges are then chosen taking this retail tariff as given. Nevertheless, it seems fair to say that in practice Ramsey pricing principles are not often heeded for regulated retail tariffs, and access charges are left to correct for the various resulting retail distortions. One possible reason for this is that retail tariffs are much more "visible" than access charges, and decision makers are more susceptible to public and political pressure when they choose these. (This is not to deny that firms in the industry are a powerful lobbying force when it comes to influencing the access charging regime.)

A common argument against the use of Ramsey prices is that they are very informationally demanding, and that compared to, say, cost-based prices, they require knowledge about demand elasticities and so on, which the regulator simply does not have. While it is broadly true that regulatory bodies do not in fact possess this kind of detailed market information, this is not to say that with further effort they could not obtain reasonable estimates of these data. Alternatively, ways could perhaps be found to delegate pricing decisions to the firm, which will most likely be much better informed than the regulator about the market in which it operates. The global price cap proposed in Section 4.7 of Laffont and Tirole (2000) is one way this might be done.

Another, more dubious, argument is that Ramsey-style access charges are "discriminatory", in the sense that the charge for a given access service will depend on the use to which this service is put. Our formula (32), for instance, shows that, as well as the cost of providing access, the access charge should depend on (i) the incumbent's (endogenous) price-cost margin in the relevant

retail market, (ii) the displacement ratio σ, which takes account of demand-side substitution possibilities as well as supply-side bypass possibilities, and (iii) the elasticity of demand for access. Each of these factors will depend on the particular use made of the access service. But the point is that this is *desirable*: Ramsey charges are the "least-bad" departures from cost-based charges, and access charges should be higher when used for those services that do not generate large welfare losses when significant price-cost markups are imposed.

It is fair to say the Ramsey approach does not enjoy the same passionate level of support or disparagement from economists as do the other two policies. One might speculate that this is because the policy is not strongly supported by either incumbents or entrants in the industry. (Very roughly, cost-based access pricing is supported by entrants, whereas the ECPR, in its simple forms at least, is supported by incumbents.) In regulatory hearings around the world, these firms find economists to support their respective cases, but there is no well-funded entity that argues strongly for Ramsey pricing.

2.8.3. The ECPR

The ECPR must be one of the more misunderstood formulas in industrial economics—see Section 2.3.1 for a variety of interpretations. The analysis in this section has argued, broadly speaking, that with our preferred version of the rule as given by (17), the rule is valid when (a) the incumbent's retail tariff is fixed in advance and is not affected by any actions of the rivals, and (b) when other instruments, such as an output tax levied on rivals, are not available to correct for the incumbent's retail market distortions. In particular, the rule has little relevance when the incumbent is free to choose its retail tariff, and in some cases it does nothing to constrain incumbent monopoly power—see Section 2.6. (However, with *partial* deregulation the rule tends to perform better.)

In the guise of the margin rule (20) it has the virtue of simplicity and being informationally undemanding (at least when compared to the Ramsey approach). All that needs to be known is the incumbent's retail price and its avoidable costs in the retail segment. However, this margin rule is simply not appropriate except in a few special cases, such as the one discussed in Section 2.3.2 for instance. If the margin rule is used as a basis for policy then inefficiencies will result in situations where (i) the rival's product does not substitute one-for-one for the incumbent's, or (ii) where rivals have some ability to bypass the incumbent's access service.

On the other hand, our preferred version of the ECPR, as represented by (17), is, outside these same special cases, informationally demanding, and various awkward elasticities appear in the expressions for opportunity cost. This suggests that the apparent contrast between the "simple" ECPR approach and the "complex" and informationally demanding Ramsey approach may just be a misleading artefact of simple examples with extreme elasticities. Indeed, the two approaches can be quite similar. This point is illustrated by the close relationship

between the ECPR and the Ramsey approach indicated in our formulas (31) and (32). Thus the difference between the two approaches simply has to do with the social cost of public funds. There would be a second source of difference related to demand cross-elasticities, and so on, if the ECPR were identified with the simple margin rule, but to do so would be to use the wrong measure of opportunity cost.

It is clear that for both ECPR-style pricing and Ramsey-style pricing there is a formidable need for information about demand and supply elasticities. Estimation of the relevant elasticities will inevitably be imperfect, and estimation errors imply efficiency losses. The extent of those losses can, however, be diminished if the incumbent's profit (or loss) on retail sales can be reduced—for example, by allowing cost-reflective tariff rebalancing.

3. Competitive bottlenecks

This section is a bridge between the two main sections of the chapter, and discusses the case of what might be termed "competitive bottlenecks". The discussion is framed in terms of two applications: call termination on mobile networks (in Section 3.1) and access pricing for competing internet networks (Section 3.2). The basic underlying features of these kinds of markets include: (i) there are several networks competing vigorously for the same pool of subscribers, and (ii) even though networks compete vigorously *for* subscribers, they often have a monopoly position in providing communications services *to* their subscribers. In such markets it is undesirable to leave network access arrangements entirely to the discretion of individual networks: there is, at the least, a role for coordination between networks, and in many cases a role for regulatory intervention to set access charges at the appropriate level.

3.1. Mobile call termination

First consider the market for mobile telephony[57]. Here the features to emphasize are that each subscriber has only one channel of communication, so that there is no "dual sourcing" of mobile telephony, and that tariff arrangements are such that the caller pays for the whole cost of the call. Given that these conditions are satisfied then, when a subscriber signs up with a network, that network has a monopoly over delivering calls to the subscriber, and it can extract monopoly profits from the callers to this subscriber. Even if the market for subscribers is intense, so that overall profits are eliminated in the sector, these monopoly profits—and the consequent deadweight losses—persist and can be used to finance subsidized retail tariffs offered to attract subscribers.

[57] This section is based on Armstrong (1997). See also Hausman (2002) and Wright (2002) for further analysis.

This topic is analyzed first in a simple model that makes the main points most clearly, and then various generalizations are discussed. Suppose, then, that there are two telecommunications sectors: fixed and mobile. In this benchmark model the latter is assumed to be competitive, and all mobile charges except possibly for call termination are unregulated. The simplifying assumptions in this section are:

Assumption 1: All calls made from mobile networks are terminated on the fixed sector;
Assumption 2: Mobile subscribers gain no utility from receiving calls;
Assumption 3: Mobile subscribers do not care about the welfare of the people who call them;
Assumption 4: Mobile subscribers do not pay anything for receiving calls made to them;
Assumption 5: The mobile sector is perfectly competitive.

We make Assumption 1 so that the charge for call termination on mobile networks does not affect the cost for making calls from mobile networks[58].

The cost structure of a mobile network is assumed to consist of a fixed cost k per subscriber, a constant marginal cost c^O for outbound calls (by Assumption 1, these are always made to the fixed sector) and a constant marginal cost c^T for terminating calls from the fixed sector. (All mobile networks are assumed to be identical.) The outbound cost c^O includes any termination payments made to the fixed sector, which are assumed to be constant in this analysis. The fixed per-subscriber cost k is the cost of a mobile handset together with any other costs associated with managing subscribers (such as billing costs). In sum, if a subscriber receives Q calls and makes q calls, a network's costs for that subscriber are $c^T Q + c^O q + k$.

Fixed network operators are required to pay a charge a per call to a mobile network for call termination. Suppose that with this charge the retail price for calls from the fixed network to the mobile network is $P(a)$. (If different mobile networks choose different termination charges, then this is reflected in different call charges from the fixed sector.) With a given fixed-to-mobile price P, suppose each subscriber on a mobile network receives $Q(P)$ calls from the fixed sector. Suppose a mobile network offers its subscribers a fixed charge f and a per-call charge p for making calls. Suppose that once a subscriber has joined a mobile network with per-call charge p, that subscriber makes a quantity $q(p)$ of outbound calls. (This is assumed for now not to

[58] If Assumption 1 does not hold then the analysis becomes closely related to the next section on "two way" access pricing. In fact, the final model in Section 4.2.4 below is closely related to the model presented here, but allows for traffic between "mobile" networks. The analysis there is simplified by the assumption that the fraction of calls that are made to other subscribers on the *same* mobile network is assumed to be negligible.

depend on the price of incoming calls P.) Then a mobile network's profit per subscriber is

$$\Pi = \underbrace{(p - c^O)q(p) + f - k}_{\text{profit from subscription}} + \underbrace{(a - c^T)Q(P(a))}_{\text{profit from termination}}. \tag{41}$$

From Assumption 5 equilibrium in the mobile sector is such that (i) operators' overall profits Π are driven down to zero, and (ii) subscriber utility is maximized subject to a network's break-even constraint. The consumer surplus of each mobile subscriber is $v(p) - f$, where $v'(p) \equiv -q(p)$. From Assumptions 2, 3 and 4 this competitive equilibrium involves marginal cost pricing for outbound calls, so that $p = c^O$, and the fixed subscriber charge recovers any profit shortfall, so that from (41) we have

$$f = k - (a - c^T)Q(P(a)). \tag{42}$$

This implies that the choice of termination charge a has no effect on the charge for outbound calls p, but only on the fixed charge f. Thus, if a is above the cost of call termination then $(a - c^T)Q(P(a)) > 0$ and a mobile operator subsidizes the retail charge for network connection in the sense that $f < k$. Thus in this particular model handset subsidies, which are a feature of many mobile markets, are a direct result of charging for call termination above cost[59].

3.1.1. Unregulated termination charges

Suppose first that mobile networks are free to choose all their charges, including termination charges. In this model it is clear what will happen: competition will drive overall profits to zero, but networks in equilibrium must maximize their profits from call termination in (41). Therefore, a is chosen to maximize

$$(a - c^T)Q(P(a)). \tag{43}$$

Call this unregulated termination charge a^{mon} for future reference. Thus, even with perfect competition *for* mobile subscribers, there is no competition for providing access *to* mobile subscribers. In particular, call termination is not

[59] In some ways this mechanism is similar to that which functions in markets with "switching costs"—see Klemperer (1995). When consumers incur switching costs for changing their supplier for a given service, a supplier has a degree of market power once a consumer has been signed up. Therefore, in competitive markets, suppliers will offer subsidized initial deals, these being funded out of the future profits generated once consumers become locked in. The difference between the two cases is that with "switching costs" it is consumer's "future self" who is exploited, whereas in the "competitive bottleneck" case it is *other* people who suffer.

charged for at the correct rate, and there is a role for regulatory intervention[60]. The socially optimal access charge is analyzed in the next section.

Before that, however, it is convenient to discuss the case where subscribers *do* care about receiving calls, since this is sometimes used as an argument for why call termination is not necessarily a "bottleneck" in the sector. So suppose that Assumption 2 does not hold, and that each mobile subscriber gains utility b per call from the fixed sector, in addition to the direct utility from making calls of $v(p) - f$. Therefore, with the termination charge a total subscriber utility is $v(p) + bQ(P(a)) - f$. Maximizing this expression subject to profits in (41) being non-negative implies that the equilibrium unregulated termination charge maximizes

$$(a - [c^T - b])Q(P(a)). \tag{44}$$

Thus, comparing (43) with (44) we see that the effect of introducing call externalities is indeed to reduce the equilibrium unregulated termination charge. (In fact, the effect is precisely as if the cost of termination is reduced by the factor b.) In the next section we discuss, among other things, whether this is a good argument for not regulating such charges.

3.1.2. Regulated termination charges

For now, return to the case where subscribers do not value incoming calls. To derive the optimal termination charge we initially make three further simplifying assumptions:

Assumption 6: The number of mobile subscribers is not affected by tariffs in the mobile market (over the relevant range of tariffs);
Assumption 7: The price of calls from the fixed to the mobile sector is equal to perceived marginal cost;
Assumption 8: Changes in the fixed-to-mobile call price have no effect on the demand for any other services offered by the fixed sector.

Assumption 7 might hold either because of competition in the fixed sector or because of optimal regulation in the fixed sector. If the marginal cost on the fixed network for making calls to the point of interconnect with mobile networks is C, then the total perceived marginal cost of calling a mobile subscriber is $C + a$. Assumption 7 therefore implies that $P(a) \equiv C + a$.

[60] In some situations this monopoly charge could be extremely high. For instance, Gans and King (2000) discuss the case where the price of calls to mobiles is a function of the *average* cost of call termination across all mobile networks. (One reason for this might be that fixed subscribers are ignorant about the mobile network they are trying to call and about the associated tariff. They therefore base their calling decisions on the average charges for calls to mobiles.) In this case a small mobile network's charge has a negligible effect on the average call termination charge, with the result that its monopoly access charge will be very high indeed.

Consumer surplus per mobile subscriber in the market for calls *to* mobile subscribers, when that price is P, is $V(P)$, where $V'(P) \equiv -Q(P)$. Therefore, from (42) and Assumptions 6, 7 and 8, total welfare per mobile subscriber when the termination charge is a is

$$W = \underbrace{V(C+a)}_{\text{utility of callers to mobiles}} + \underbrace{v(c^O) + (a - c^T)Q(C+a) - k}_{\text{utility of mobile subscribers}}. \tag{45}$$

(Profit in both sectors is zero, and Assumption 8 implies that the fixed-to-mobile price has no effect on other profits in the fixed sector. Assumption 6 implies that welfare per mobile subscriber is an accurate and well-defined measure of total welfare.) This expression is maximized by setting $a = c^T$, so that there should be marginal cost pricing of call termination under our assumptions. This in turn implies that $f = k$ in equilibrium, and there are no handset or other subsidies for mobile network connection at the optimum. Although mobile subscribers certainly benefit from high termination charges—since their network connection is subsidized as a result—this benefit is more than outweighed by the costs this imposes on their callers.

Although this simple model captures some important features of the market, it ignores a number of important aspects. The next section adapts the model to discuss the determination of access charges in more complex environments.

3.1.3. Extensions

Price of fixed-to-mobile calls not equal to cost. Suppose that Assumption 7 does not hold, and that the price of calls to mobile subscribers is above the perceived marginal cost. This might be, for instance, because regulation in the fixed sector required that call charges be above cost in order to fund social obligations, or because this call charge is unregulated. Then, with no other modifications to the above benchmark model, it is optimal to have the termination charge be such that the price of calls to mobile subscribers equal to the *actual* cost of making such calls, i.e. so that $P(a) = C + c^T$. When $P(a) > C + a$ this obviously implies that $a < c^T$, and so it is optimal to subsidize call termination on mobile networks in order to counteract the price-cost mark-up in the fixed-to-mobile call market.

Allowing for network externalities.[61] Suppose next that Assumption 6 does not hold, so that the number of mobile subscribers is elastic[62]. Specifically, if $U = v(p) - f$ is the net surplus offered by the mobile networks, suppose that the total consumer surplus of the mobile subscribers is $\Phi(U)$ and the number of

[61] In this discussion we assume that the call termination charge on the fixed network is equal to cost, in order to ignore the effect on fixed network profits caused by changes in mobile penetration.

[62] More generally, much of the analysis in Willig (1979) is to do with tariff design in the presence of network and call externalities.

mobile subscribers is $N(U) = \Phi'(U)$, where $N(\cdot)$ is an increasing function. (We assume that potential mobile subscribers differ only in their willingness to subscribe, and not in the number of calls they make once they do subscribe.)

Callers on the fixed network benefit from higher mobile subscription, as there are then more people to call on the mobile network. As above, suppose that $Q(P)$ is the number of fixed-to-mobile calls per mobile subscriber, which is assumed not to depend on the number of mobile subscribers. The fact that all potential mobile subscribers are equally valuable to fixed subscribers, which is what this constant calls per mobile subscriber assumption implies, means that the "network externality" is linear in the number of mobile subscribers. In particular, consumer surplus in the fixed sector is $NV(P)$ when there are N mobile subscribers and the fixed-to-mobile call charge is P.

As before, in equilibrium overall mobile network profits are zero and mobile subscriber utility is maximized given this zero profit constraint. This again implies that $p = c^O$ and f is given by (42), so that $U(a) = v(c^O) + (a - c^T)Q(C + a) - k$. For values of a less than a^{mon} in Section 3.1.1, $U(a)$ is increasing in a since a higher termination charge increases the profits from call termination, which then translates into a lower fixed charge f. This in turn translates into higher subscriber numbers for the mobile sector. Total welfare, which was previously given by (45), is now

$$W = \underbrace{N(U(a))V(C+a)}_{\text{utility of callers to mobiles}} + \underbrace{\Phi(U(a))}_{\text{utility of mobile subscribers}}.$$

Maximizing this with respect to a gives the first-order condition

$$a = c^T + \frac{N'(U(a))U'(a)V(C+a)}{-NQ'(C+a)} \geqslant c^T.$$

Some intuition for this opaque formula can be given as follows. In general, there is a deadweight loss associated with subsidizing mobile subscribers. That is to say, an extra \$1 subsidy to each mobile subscriber costs the fixed subscribers $\$(1 + \Lambda)$ per mobile subscriber in lost consumer surplus, $\Lambda \geq 0$ is the deadweight loss paramerter. Since the subsidy paid to each mobile subscriber is $\pi_T(a) \equiv (a - c^T)Q(C + a)$, to boost the subsidy by \$1 means that a must increase by $1/(\pi'_T(a))$. Since $\frac{d}{da}V(P(a)) \equiv -Q(C + a)$, this costs the fixed users $\$Q/\pi'_T$ per mobile subscriber. In sum,

$$\Lambda(a) = \frac{Q(C+a)}{\pi'_T(a)} - 1 = \frac{-(a - c^T)Q'(C+a)}{\pi'_T(a)}$$

For $a > c^T$ (and $a < a^{mom}$), $\Lambda(a)$ is strictly positive (i.e., there are deadweight losses) unless the demand for calls to mobiles is perfectly inelastic.

With this notation, the above opaque formula for a reduces to

$$N'V = N\Lambda.$$

The left-hand side of this expression is the 'external' benefit to fixed subscribers of a marginal increase in the subsidy to mobile subscribers, while the right-hand side represents the total deadweight loss incurred. Notice that if $N' \equiv 0$, so that mobile penetration does not respond to subsidies, then the formula implies that $\Lambda(a) = 0$ at the optimum, which then entails marginal cost pricing of termination as we derived previously in Section 3.1.2. At the other extreme, if $Q' \equiv 0$, then $\Lambda(a) \equiv 0$ and there are no deadweight losses associated with transferring funds from fixed to mobile subscribers. In this case the formula implies that mobile penetration should be increased to the point where $N' = 0$. Since there is no welfare cost associated with subsidizing mobile subscription when $Q' = 0$, sufficiently high subsidies should be offered that every potential mobile subscriber actually subscribes.

In general, however, the network externality effect implies that $\Lambda(a) > 0$, which implies that it is optimal to set the termination charge above cost.

The reason for this is clear: a higher termination charge raises the equilibrium mobile subscriber utility via handset subsidies and the like, this in turn increases mobile subscription, which in turn raises the utility of fixed network subscribers because of the network externality effect.

However, even though the presence of network externalities gives a reason for pricing call termination above cost, the optimal termination charge is still lower than the unregulated charge a^{mon} in Section 3.1.1 above[63]. Therefore, the realistic assumption that network externalities are important in mobile telephony does not provide a good argument for deregulating mobile call termination charges[64].

Allowing for call externalities. As was earlier mentioned in Section 3.1.1, suppose a subscriber gains utility b from receiving each call. Given Assumption 7 welfare per mobile subscriber is modified from (45) to

$$W = V(C + a) + v(c^O) + bQ(C + a) + (a - c^T)Q(C + a) - k.$$

The optimal termination charge therefore satisfies the first-order condition

$$a = c^T - b < c^T. \tag{46}$$

[63] The regulator would not choose a termination charge above a^{mon} since both mobile subscribers and callers to mobiles would be made better off by reducing a down to a^{mon}. Also, welfare is increased by reducing a at least a little below a^{mon} since this has only a second-order effect on U (and hence on mobile subscriber numbers) and yet yields a first-order benefit for callers to mobile subscribers.

[64] In any event, it is possible that there exist superior ways to fund the (desirable) subsidy to mobile subscribers instead of imposing a price-cost margin on the price of calls from fixed-to-mobile networks. (Perhaps a small tax on the whole industry would be less distortionary than a big tax on one sector?)

Therefore, if mobile subscribers derive a benefit from incoming calls, then the regulator should set the termination charge *below* cost in order to encourage calls from the fixed sector.

This result means that the presence of call externalities is *not* a good reason to de-regulate call termination. In Section 3.1.1 we showed that call externalities did cause networks to reduce their termination charges, in order to stimulate demand for calls to their subscribers. However, we have also seen that call externalities act to reduce the socially optimal charge. Comparing (46) with (44) we see that unregulated networks still price termination in excess of the socially optimal level.

Allowing subscribers to care about their callers. This extension provides a better argument for de-regulating call termination. Suppose next that Assumption 3 does not hold, and for simplicity suppose that mobile subscribers internalize the entire welfare of those who call them when they choose their mobile network. (For instance, if the regular callers are family members or close friends, this might be a good approximation.) Mobile networks will still compete away all profits, and total welfare of mobile subscribers—which now includes that of callers on the fixed network—is maximized subject to this break-even constraint. Suppose that mobile firms are free to choose their own termination charge. If Assumption 7 continues to hold, then it is clear that marginal cost pricing will be the equilibrium outcome: $p = c^O$, $f = k$ and $a = c^T$. Therefore, if mobile subscribers and those who call them are mostly closely-knit groups, the need for the control of mobile call termination in a competitive market is much reduced.

Allowing for charging for incoming calls. In some countries, such as the United States, the arrangement is that calls from the fixed to the mobile sector are charged at regular fixed network tariffs, and the recipient of the call also makes a payment per call to cover the additional cost of terminating on mobile networks[65].

To be concrete, suppose that it is now the mobile subscriber who pays for call termination, not the caller. Then Assumption 7 states that with this arrangement we have $P = C$. If a mobile network charges a per call to *its* subscribers for receiving incoming calls, then its overall profit from its subscribers is modified from (41) to be

$$\Pi = (p - c^O)q(p) + f - k + (a - c^T)Q(C).$$

[65] This arrangement is often due to country-specific numbering issues, such as callers being unable to distinguish mobile and fixed telephone numbers due to the former having no distinct number ranges. See Kim and Lim (2001), Hermalin and Katz (2001) and Jeon, Laffont and Tirole (2001) for analysis of the general issue of reception charges in telecommunications.

If mobile subscribers have no choice but to accept incoming calls, being charged for these calls is exactly equivalent to increasing the fixed charge[66]. (Recall that all mobile subscribers receive the same, known, number of calls in this model.) Therefore, the balance between the fixed charge f and the incoming call charge a is not determined here. Equilibrium here involves $p = c^O$, as usual, and $f + (a - c^T)Q(C) = k$, so that all other charges cover the mobile network's fixed costs.

Clearly the distortion is no longer that mobile operators exploit their market power over delivering calls to their subscribers—that problem has been eliminated in competitive markets by making mobile subscribers themselves pay for call termination—but that callers on the fixed networks now pay too little. (The price P should, in the absence of call externalities, be equal to the total cost $C + c^T$ rather than just C.) Of course, it is just possible that one market failure will cancel out another, and if there is a call externality b that is precisely equal to the termination cost c^T, this receiver-pays tariff arrangement might be desirable. However, the two effects are again quite different, and the presence of call externalities is not in itself an argument for making call recipients rather than callers cover the cost of call termination.

Allowing for partial substitution between fixed and mobile sectors. Finally, suppose that Assumption 8 does not hold. There are several ways in which charges for calls to and from the mobile sector affect the take-up of other services provided by the fixed sector. For instance, low charges for mobile services could cause some subscribers to leave the fixed network altogether and make all their calls on their mobile network, which of course reduces the demand for fixed services. Alternatively, consider a person who subscribes to both fixed and mobile services. Since calls to this person using the two networks are substitutes, we imagine a reduction in the price of fixed-to-mobile calls would, all else equal, reduce the demand for fixed-to-fixed calls. For simplicity, we focus on the second of these two kinds of substitutability.

Suppose that, if the fixed-to-mobile call charge is P, total profits (per mobile subscriber) in the rest of the fixed sector are $\pi(P)$. (We suppose that charges for all other fixed services are not affected by policy towards the mobile sector.) Assumption 7 implies that total welfare is modified from (45) to

$$\pi(C + a) + V(C + a) + v(c^O) + (a - c^T)Q(C + a) - k.$$

(Here, $V(P)$ is consumer surplus in the fixed sector keeping all other fixed sector charges constant.) Maximizing this with respect to a implies that

$$a = c^T + \frac{\pi'(C + a)}{-Q'(C + a)}. \tag{47}$$

[66] The papers mentioned in the previous footnote allow the call recipient to cut short, or not accept, a call if the charge outweighs the benefit of being called.

Thus the sign of $(a - c^T)$ is the same as that of π'. It is plausible that π' is positive since increasing the fixed-to-mobile price will increase the demand for fixed-to-fixed services, which will usually add to profits. If this is so then it is optimal to set the termination charge above cost, in order to stimulate demand for (profitable) fixed-to-fixed services. Expression (47) is a variant of the ECPR formula (17). In order to provide the marginal unit of access to the fixed sector, the mobile termination charge must increase by $1/Q'$. Therefore, the second term in the above expression is the lost profit in the *fixed* sector caused by the *mobile* sector supplying a unit of access.

3.2. Access charges for the internet

A market that in several ways is quite similar to that of the above model of mobile telephony is the internet. A simple model of the internet has two classes of agent: web-site providers (who provide information and content of various kinds) and consumers (who wish to obtain content provided on the web-sites)[67]. In this simple model, all information flows are "one way", from web-sites to consumers. Consumers obtain utility from viewing content on the web-sites, and web-site providers obtain utility from consumers visiting their web-sites. There are several means by which web-site providers might obtain utility from consumers. A "commercial" web-site might sell content to visitors, either electronically (such as an economics journal being published electronically) or acting as an on-line retailer for other kinds of products (such as *Amazon.com*). Alternatively, a web-site provider might gain utility even if there is no direct payment from the visitor. For instance, it might make money from providing advertisements to its visitors, from providing useful information about a firm's products (which are then purchased by conventional means), from saving money on conventional postage when content is downloaded from the site, or it might simply obtain utility from knowing people have visited the site. For simplicity, in this section we consider the second kind of web-site, and suppose that web-site providers do not charge consumers directly for entry to the site.

Consider first a benchmark model where the number of consumers and the number of web-sites is exogenously fixed (provided that some reservation level of utility is obtained by the two groups), with the number in each group being normalized to 1. Suppose each consumer obtains utility u from visiting each web-site, and each web-site obtains utility $\hat{u}$ from each consumer who visits it. There are a number of identical internet networks operating in a perfectly competitive market. The total cost of carrying a unit of communication from a web-site to a consumer is $c^O + c^T$, where c^O is the cost of *originating* communication from the

[67] The following discussion is based on Laffont, Marcus, Rey, and Tirole (2001). See also chapter 7 of Laffont and Tirole (2000) and Little and Wright (2000).

web-site and c^T is the cost of *terminating* the communication to the consumer. If a consumer connected to one network visits a web-site hosted on a rival network, the host network incurs the cost c^O while the terminating network incurs the cost c^T.

Since the act of a consumer visiting a web-site benefits both parties—it is similar to the above model of mobile telephony with the extension to allow for call externalities—it is not obvious whether the terminating network should be paid by the originating network or *vice versa*. Here we adopt the convention that the web-site's network pays the access charge a to the consumer's network for delivering the communication. If a is negative, however, this arrangement implies that the web-site's network is paid for providing the service of delivering the communication to the consumer's network. In this section we assume that every network sets the same termination charge a. Turning to the retail side, suppose network i charges its consumers p_i each time they access any web-site and charges its web-sites $\hat{p}_i$ each time any consumer visits them.

Putting all of this together implies that if network i attracts n_i consumers and $\hat{n}_i$ web-sites, its total profits are

$$\Pi_i = \underbrace{\hat{n}_i(\hat{p}_i - c^O - n_i c^T - (1 - n_i)a)}_{\text{profits from web-sites}} + \underbrace{n_i(1 - \hat{n}_i)(a - c^T)}_{\text{profits from termination}} + \underbrace{n_i p_i}_{\text{profits from reception charges}}.$$

(Note that the cost allocation involved in this decomposition is quite arbitrary, as we have loaded all of the costs involved in "completing calls" onto the web-site segment of market.) This can be simplified to

$$\Pi_i = n_i(p_i + a - c^T) + \hat{n}_i(\hat{p}_i - a - c^O). \tag{48}$$

Therefore, the profits of a network can be decomposed into those generated by selling services to web-site providers and those generated by its services to consumers. The effective marginal cost of providing services to consumers is $c^T - a$, while the effective marginal cost of providing services to web-sites is $c^O + a$.

Given the assumption of perfect competition between networks, equilibrium prices are driven down to the associated marginal costs:

$$p_i = c^T - a; \quad \hat{p}_i = c^O + a \tag{49}$$

and network profits are zero[68]. Thus, the choice of regulated access charge affects the balance of retail charges offered to consumers and to web-site providers in equilibrium. If there is no access charge for interconnection, so that $a = 0$ and a "bill and keep" system is used, then consumers are charged the cost of terminating communications to them, while web-sites are charged the cost of originating communication. By contrast, if access is charged at termination cost, so that $a = c^T$, then consumers pay nothing for using the internet while web-sites pay the full cost of providing communication. Alternatively, if the originating network can recover its costs, so that $a = -c^O$, then the reverse holds.

Notice in particular that a network's charging strategy in (49) involves setting its retail prices as though all terminating traffic came from rival networks and as though all originating traffic was to be terminated on rival networks. For instance, a web-site connected to a given network will have a fraction of visits from consumers connected to the same network, and yet the origination charge is $c^O + a$ which is the cost of sending communications to a rival network. Laffont, Marcus, Rey, and Tirole (2001) call this result the "off-net-cost pricing principle"[69].

While this simple model with inelastic demand is suitable for demonstrating how the access charge feeds through into retail charges in equilibrium, it is incapable of analyzing the normative issue of the optimal level of the access charge. (As long as all consumers and web-site providers are served, welfare is not affected by the balance between the two retail charges.) Therefore, we next extend the model to allow for some responses to prices. One way to do this is to suppose that there is elastic consumer demand for visiting web-sites (but the number of web-sites remains fixed). If the price for receiving content is p, suppose that each consumer makes $q(p)$ visits to each web-site. Otherwise, everything is as described above. Then, after some manipulation, (48) becomes

$$\Pi_i = n_i q(p_i)(p_i + a - c^T) + \hat{n}_i[n_i q(p_i) + (1 - n_i)q(\bar{p})](\hat{p}_i - a - c^O),$$

where $\bar{p}$ is the (average) retail price for receiving content for consumers on the rival networks. Unlike the previous case with inelastic consumer demand, here a

[68] For this result to be valid, these candidate prices must not exceed the agents' reservation utilities, so that the access charge should satisfy $\hat{u} - c^O > a > c^T - u$.

[69] This insight implies that the market works just as if there are two kinds of network, one kind that just caters for consumers and one kind that just caters for web-sites. There is one-way traffic from the latter set of networks to the former. Viewed from this perspective, this market is very similar to the model of mobile telephony in the previous section, with call traffic from the fixed networks to the mobile networks. (The fact that mobile networks also called the fixed network played no role in the previous analysis, since the termination charge on the fixed networks, and hence the quantity of calls to the fixed network, was exogenously fixed.)

network's profits do not neatly decompose into profits from consumers and profits from web-sites. Nevertheless, the off-net-cost pricing principle still holds, and equilibrium prices are given by (49) above[70]. Equilibrium profits are zero as before. However, unlike the inelastic case where the level of the access charge had no effect on welfare, here there is a welfare effect. Assuming everyone is served, welfare per web-site with the reception charge p is

$$v(p) + [p + \hat{u} - c^O - c^T]q(p).$$

(Recall that a web-site gains utility $\hat{u}$ from each visit. Here $v(p)$ is the consumer surplus function associated with $q(p)$, so that $v'(p) = -q(p)$.) Maximizing this implies that the ideal reception price is

$$p^* = c^O + c^T - \hat{u},$$

and so from (49) this implies that the optimal access charge that implements this outcome is

$$a = \hat{u} - c^O. \tag{50}$$

This access charge also implements the origination charge $\hat{p} = \hat{u}$, which means that web-site providers are left with none of the gains from trade in equilibrium. This is to be expected: Ramsey principles suggest that any service inelastically supplied should bear the burden of price-cost markups. If the supply of web-sites was also elastic, then the optimal access charge will have to trade-off the efficiency losses of both sides of the market[71]. However, the formula (50) does not necessarily imply that the access charge is high, or even positive: if $\hat{u}$ is small, so that most of the benefits of communication are on the consumer side, then the access charge will be negative, i.e. the originating network is paid for supplying valuable information to consumers[72].

[70] For instance, if the rival networks offer a reception price $\bar{p}$ which is greater than the effective marginal cost $c^T - a$, then it pays network i to undercut this price slightly and to take the entire population of consumers.

[71] See Laffont, Marcus, Rey, and Tirole (2001) for this analysis, and also for several further extensions to this basic model (including allowing for imperfect competition between networks, and for differential charging according to whether traffic is "on-net" or "off-net").

[72] As well as this Ramsey analysis, Laffont, Marcus, Rey, and Tirole (2001) discuss how networks would cooperatively choose the access charge a to maximize profits. (They extend this competitive model to allow for imperfect competition.) They show that there is no reason to expect that networks will agree to charge the socially optimal charge, and so there is some scope for regulatory intervention in this market.

This brief discussion of some issues concerning access pricing between internet networks is similar in some ways to the earlier analysis of the mobile telephony market. In particular, even though there is effective competition for both consumers and web-site providers, this does not justify a *laissez-faire* policy towards interconnection arrangements between networks. Perhaps the main difference between the mobile model and the internet model is that in the latter there are *two* bottlenecks: (i) once a consumer has signed up with a network, that network has a monopoly over providing communication to that consumer (similar to the case of mobile subscribers above), and (ii) once a web-site has signed up with a network, that network has a monopoly for originating communication from that web-site[73].

4. Two way access pricing and network interconnection

In this section of the chapter we discuss the important topic of two-way network interconnection. In contrast to the scenarios outlined in previous sections, here all firms in the market must negotiate with each other to gain access to each other's subscribers[74]. Because this analysis can quickly become complicated, we look first at the case of fixed, captive subscriber bases, which is naturally illustrated by the case of international call termination. In subsequent sections we allow these market shares to be determined endogenously by the competitive process.

4.1. Fixed subscriber bases: International call termination

Consider two countries, A and B[75]. Suppose that the cost of originating a call in country i (to be terminated in the other country) is c_i^O and the cost of terminating

[73] This latter effect is also present in the mobile model—once a subscriber has joined a particular mobile network, that network must originate all calls from the subscriber—but is unimportant when there is effective competition. The difference is that in the mobile model people on the fixed network were assumed not to obtain any benefit from being called by mobile subscribers and, more importantly, they were not charged for receiving calls.

[74] In fact the previous model of internet interconnection also had this feature, in that networks typically would have both consumers and web-sites as customers, and so networks would need access to each others' consumers to deliver communications originating on their own networks. However, in that model originators and recipients of communications were disjoint groups, and information flows were always in one direction. This feature greatly simplifies the analysis.

[75] See Hakim and Lu (1993), Carter and Wright (1994), Cave and Donnelly (1996), Yun, Choi and Ahn (1997), Section 6 of Laffont, Rey, and Tirole (1998a), Domon and Kazuharu (1999), Wright (1999), and Box 5.1 in Laffont and Tirole (2000) for more analysis of the economics of international settlements in telecommunications. At a deeper level, this analysis is closely related to the problem of negotiating trade tariffs/subsidies between two large countries—for instance, see Mayer (1981) for a classic treatment.

a call in country i (originating in the other country) is c_i^T. Let the total cost of making a call from i to j be $c_i = c_i^O + c_j^T$. The price of a call from country i to country j is p_i. The demand for calls from i to j is $x_i(p_i)$[76]. Let π_i be defined by $\pi_i(p_i) = (p_i - c_i)x_i(p_i)$, and this is the profit function in country i if call termination happens to be priced at marginal cost. Consumer surplus in i is $v_i(p_i)$, where $v_i' = -x_i$. (We assume that the prices in this international market do not affect the demand for other telecommunications services supplied in the two countries.) Total (world) surplus due to the call traffic between the countries is therefore

$$v_A(p_A) + \pi_A(p_A) + v_B(p_B) + \pi_B(p_B),$$

which is maximized by setting prices equal to the actual marginal costs: $p_i = c_i$. Let the call termination charge in country i be a_i. Then the profits of country i due to this international market are

$$\Pi_i = \underbrace{(p_i - c_i^O - a_j)x_i(p_i)}_{\text{profits from call origination}} + \underbrace{(a_i - c_i^T)x_j(p_j)}_{\text{profits from call termination}} . \tag{51}$$

Assume that the move order is for termination charges a_i to be chosen first and then, taking these as given, countries choose their retail prices p_i non-cooperatively[77]. Therefore, given the pair of access charges (a_A, a_B), each country i chooses its retail price p_i to maximise its welfare, taking the other's retail price as given.

Suppose first that retail price regulation (or competition) in each country is such that, given the foreign country's termination charge, the call charge is equal to the perceived marginal cost of the call:

$$p_i = c_i^O + a_j. \tag{52}$$

(This is similar to Assumption 7 in Section 3.1.2.) Clearly, if

$$a_i = c_i^T \quad \text{for } i = A, B \tag{53}$$

[76] For simplicity, assume there are no cross-price effects for calls in the two directions. See Acton and Vogelsang (1992) for evidence that cross-price effects are not significant.

[77] The reverse ordering (or assuming all prices are chosen simultaneously) does not make sense since if the two retail prices are fixed then the quantity of calls which country i must terminate, which is $x_j(p_j)$, is also fixed. (we assume a country cannot refuse to terminate calls at the specified price.) Country i can then set an arbitrarily high termination charge a_i and make arbitrarily high profits, and no equilibrium can exist with the alternative move order.

then both countries will set the ideal prices $p_i = c_i$. Therefore, cost-based termination charges induce the best outcome from the point of view of overall welfare, at least provided national regulators act in the way described above. Given the one-to-one relationship between a_i and p_j in (52) we can think of each country as choosing the *other* country's retail price for calls. Written in this way profits in (51) become

$$\Pi_i = (p_j - c_j)x_j(p_j) = \pi_j(p_j).$$

(There are no profits from call origination.) Welfare in country i is therefore

$$w_i = v_i(p_i) + \pi_j(p_j) \tag{54}$$

where country i chooses p_j and country j chooses p_i.

So far the analysis has been done assuming that regulation or competition in each country forces the price of international calls to be equal to the perceived cost of making such calls. International calls have historically been priced substantially above the associated costs—even taking into account the existing (high) international call termination payments—and so it is worthwhile to extend this analysis to allow for the possibility that countries might wish to use profits from international calls for other, perhaps socially useful, purposes. One way to model this is to suppose that country i receives benefits of $1 + \lambda_i > 1$ for each unit of profit it makes in the international sector, from both retail and call termination sources. (The previous analysis assumed that $\lambda_i = 0$.) In this case welfare in country i is given by

$$w_i = v_i(p_i) + (1 + \lambda_i)\{(p_i - c_i^O - a_j)x_i(p_i) + (a_i - c_i^T)x_j(p_j)\}. \tag{55}$$

Since profits are now more valuable, a country will wish to set its outbound price above its perceived marginal cost. Indeed, maximizing the above expression with respect to p_i given a_j implies that marginal cost pricing is replaced by the Ramsey formula

$$\frac{p_i - c_i^O - a_j}{p_i} = \frac{\lambda_i}{1 + \lambda_i}\frac{1}{\eta_i}, \tag{56}$$

where η_i is the elasticity of demand for calls from i to j. In this case we have $p_i > c_i^O + a_j$, although there is still a one-to-one, increasing relationship between a_j and p_i. A country with a higher "social cost of public funds" parameter λ_i or a less elastic demand for calls will choose a higher price-cost markup for its international calls. If we write

$$\phi_i(a_j) = \max_{p_i} : \{v_i(p_i) + (1 + \lambda_i)(p_i - c_i^O - a_j)x_i(p_i)\}$$

to be the maximum welfare in country i due to outbound calls, then the envelope theorem implies that

$$\phi_i'(a_j) = -(1 + \lambda_i)\hat{x}_i(a_j) \tag{57}$$

where $\hat{x}_i(a_j)$ is the optimal number of calls given the overseas country's termination charge a_j. Here, $\hat{x}_i$ is decreasing in a_j. Also, a higher value for λ_i translates into a higher value for $(p_i - c_i^O - a_j)x_i(p_i)$, which in turn induces a smaller value for $\hat{x}_i$.

4.1.1. Non-cooperative determination of termination charges

Suppose now that each country chooses its termination charge—or equivalently, the overseas country's retail price—non-cooperatively. Given the pricing rule (52), country i knows that if it chooses the overseas retail price p_j it will make a profit from call termination equal to $\pi_j(p_j)$. Since the choice of termination charge does not affect its consumer surplus from originating calls (which is chosen by j), i will choose p_j to maximize profits from call termination, so that p_j is chosen by i to satisfy the usual monopoly formula

$$\frac{p_j - c_j}{p_j} = \frac{1}{\eta_j} > 0. \tag{58}$$

This implies that $a_i > c_i^T$[78]. Therefore, if countries choose their termination charges independently, each will set a charge above cost, with the result that world surplus is not maximized. Each country exploits its monopoly position in call termination with the result that, despite the first-best pricing behaviour at the national level in (52), retail prices are set as if networks were unregulated profit-maximizing monopolies in each direction. In particular, it is straightforward to see that each country will benefit if both termination charges are brought down at least a little from this non-cooperative equilibrium, since each country suffers only a second-order loss in profits from call termination, but a first-order gain in the surplus from making calls[79]. Thus we have the important

[78] More generally, if the overseas country has demand $\hat{x}_j(a_i)$ given the home country's termination charge—as illustrated by the case of Ramsey pricing mentioned above—then i will choose a_i to maximize its call termination profit $(a_i - c_i^T)\hat{x}_j(a_i)$.

[79] However, it is *not* true that it is in each country's interest individually to reduce both termination charges all the way down to cost. For instance, suppose one country has only a tiny demand for calls to the other country, so that $v_i \approx 0$. This country will then lose out if both charges are brought down to cost.

insight that non-cooperative setting of termination charges will cause the charges to be set at too high a level, due to the standard double-marginalization problem.

4.1.2. Cooperative determination of termination charges

Given the inefficiency of choosing termination charges non-cooperatively, it is natural and desirable that countries negotiate their mutual termination charges. Here we consider three kinds of negotiations with progressively less complex kinds of bargaining:

Bargaining with side-payments. If side-payments can costlessly be made between countries (and if there is no asymmetric information about costs and demands), it is natural to suppose that an efficient outcome is attainable, and termination charges will be set equal to costs: $p_i = c_i$ and $a_i = c_i^T$. How the first-best surplus is divided between the countries will depend on the details of the bargaining procedure.

Bargaining with non-reciprocal termination charges. Next, suppose that no such side-payments are possible, and the two countries simply bargain over the pair of access charges (assuming retail prices are then set as in (52)). Whatever the precise form the negotiations take, it is reasonable to suppose bargaining is (second-best) efficient in the sense that one country's welfare is maximized subject to the other's welfare being held constant, i.e. retail prices p_A and p_B must maximize w_A subject to $w_B \geqslant \bar{w}_B$ for some reservation level $\bar{w}_B$. From (54) the first-order conditions for this problem are

$$\frac{\pi_A'(p_A)}{x_A(p_A)} = \frac{x_B(p_B)}{\pi_B'(p_B)}$$

or

$$\left(1 - \frac{p_A - c_A}{p_A}\eta_A\right)\left(1 - \frac{p_B - c_B}{p_B}\eta_B\right) = 1.$$

In particular, either $p_i = c_i$ for both countries (which can be optimal only in knife-edge situations, depending on $\bar{w}_B$) or one country's call charge is above cost and the other's is below cost. Therefore, except in totally symmetric situations (when $p_i = c_i$ and $a_i = c_i^T$ is the outcome of any reasonable bargaining process), one country makes a loss on terminating calls, and the other makes a profit.

Bargaining with reciprocal termination charges. Finally, suppose that countries must choose symmetric termination charges, so that $a_A = a_B = a$, say. Given the pricing rule (52), welfare in (54) with a reciprocal termination charge a is

$$w_i = v_i(c_i^O + a) + \pi_j(c_j^O + a),$$

and so country i's ideal *reciprocal* termination charge, denoted a_i^*, is given by the expression

$$a_i^* = c_i^T + \frac{x_j(c_j^O + a_i^*) - x_i(c_i^O + a_i^*)}{-x_i'(c_j^O + a_i^*)}. \tag{59}$$

Therefore, in symmetric situations where $c_A^T = c_B^T$, $c_A^O = c_B^O$ and $x_A \equiv x_B$ the interests of countries coincide, and each is happy to agree to charge for call termination at marginal cost. In other cases, however, countries will have divergent interests. Given divergent preferences, it is natural to suppose that the equilibrium reciprocal charge will lie between the two privately preferred values, a_A^* and a_B^*, and its precise location will depend on the balance of "bargaining power" between the two countries.

Expression (59) shows that a country would like to have a reciprocal termination charge that is higher than its termination cost whenever (i) originating costs are approximately equal and the foreign country has higher demand for calls than the home country, or (ii) demand functions are approximately equal and the foreign country has lower costs for originating calls than the home country. In sum, countries with either high costs or a net inflow of calls over the relevant range of prices will prefer a higher reciprocal termination charge. In practice this means that poorer countries will, on the whole, prefer higher reciprocal termination charges than will more developed countries.

More generally, when the countries have a social cost of public funds, welfare (55) in country i with the reciprocal charge a is

$$w_i = \phi_i(a) + (1 + \lambda_i)(a - c_i^T)\hat{x}_j(a).$$

From (57), this is maximized by setting

$$a_i^* = c_i^T + \frac{\hat{x}_j(a_i^*) - \hat{x}_i(a_i^*)}{-\hat{x}_i'(a_i^*)},$$

which generalizes (59) above. This shows that another factor which causes preferences to diverge over termination charges is the social cost of public funds. As discussed above, in otherwise symmetric environments, if $\lambda_j > \lambda_i$ then $\hat{x}_j(a) < \hat{x}_i(a)$ and so country i prefers a lower reciprocal termination charge than

country j. Since poorer countries will also tend to have a higher social cost of public funds—due, for instance, to fewer sources for effective taxation—this will be yet another reason to expect that these countries will prefer higher reciprocal termination charges.

4.2. *Interconnection with competition for subscribers*

In this section we extend the analysis to allow networks to compete for subscribers, rather than taking market shares as exogenously fixed. While some of the insights from the analysis of international call termination continue to be valid in this competitive case—for instance, non-cooperative setting of termination charges will lead to inefficiently high retail prices—others will not.

4.2.1. *A general framework*

Consider the following model: there are two networks, A and B, competing for the same subscribers[80]. Each subscriber purchases all telephony services from one network or the other (or none). There is a continuum of potential subscribers, with total number normalized to 1. Subscribers are differentiated, and if they receive "utility" u_A from using network A and u_B from using network B, then network A has $n_A = s_A(u_A, u_B)$ subscribers and network B has $n_B = s_B(u_B, u_A)$ subscribers. (These utilities will be derived below.) Naturally, $s_i(u_i, u_j)$ is increasing in u_i and decreasing in u_j, since a better deal being offered by the rival causes a network's subscriber numbers to fall. If $n_A + n_B = 1$, then all potential subscribers join one or other network over the relevant range of utilities; otherwise there is only partial participation, and network externalities become an important ingredient of the analysis.

An example of the market share function s_i is obtained from the familiar Hotelling model of consumer choice. Suppose that the two firms are "located" at each end of the unit interval. A consumer's type (or location) is denoted $y \in [0, 1]$. If the two firms' utilities are u_A and u_B, then such a consumer gains utility of $u_A - wy$ if she joins network A, and utility $u_B - w(1 - y)$ if she joins network B. Here $w > 0$ is a parameter that determines how closely substitutable are the two services. Suppose consumers' types are uniformly distributed over the unit interval. Then s_i is given by

[80] This basic model is adapted from Armstrong (1998), Laffont, Rey, and Tirole (1998a) and Laffont, Rey, and Tirole (1998b). The first paper considered only linear retail prices, whereas the latter two analyzed the more relevant case of two-part tariffs as well. The last of these discussed the case where networks are permitted to condition their call charges upon the destination network. See also the closely related work of Carter and Wright (1999), who examine in more detail whether firms might not interconnect at all. See Chapter 5 in Laffont and Tirole (2000) for another overview of two-way network interconnection issues. The "competition in utility space" approach used in this section is explored in more detail in Section 2 of Armstrong and Vickers (2001).

$$n_i = s_i(u_i, u_j) = \frac{1}{2} + \frac{u_i - u_j}{2w} \tag{60}$$

(provided that $0 \leqslant n_i \leqslant 1$). This is an example where there is full subscriber participation, so that $n_A + n_B \equiv 1$ over the relevant range of utilities[81].

Depending on the context, networks could be using two-part pricing or linear pricing[82]. To keep the analysis as general as possible at this stage, suppose firm i offers its subscribers the tariff

$$T_i(x, \hat{x}) = p_i x + \hat{p}_i \hat{x} + f_i, \tag{61}$$

where x is the number of calls made to other subscribers on the same network i ("on-net calls"), $\hat{x}$ is the number of calls made to subscribers on the rival network ("off-net calls"), p_i is the marginal price for on-net calls, $\hat{p}_i$ is the marginal price for off-net calls, and f_i is the fixed charge (which is zero if linear pricing is used). Public policy (or technological limitations) may require that $p_i = \hat{p}_i$, so that networks cannot price discriminate according to the destination network. We assume that subscribers do not pay for receiving calls[83].

A subscriber is assumed to gain the same utility from calling every other subscriber, and if a subscriber faces the price p for calling some other subscriber, the former will make $x(p)$ calls to the latter. This demand function is assumed to be independent of the identities of the caller and recipient, and also of the network that each party has joined. (Later we will allow callers to differ in their demand for calls and in how many calls they receive.) Let $v(p)$ be the level of consumer surplus associated with the demand function $x(p)$, so that $v' \equiv -x$. For now, suppose that subscribers obtain no utility from receiving calls—this assumption is relaxed in Section 4.2.3. Then, given the network choices made by the other subscribers, the utility received by a subscriber if she joins network $i = A, B$ is

$$u_i = n_i v(p_i) + n_j v(\hat{p}_i) - f_i. \tag{62}$$

Thus a subscriber's utility is linear in the number of subscribers of each network, i.e. network externalities take the same linear form as we used in the mobile telephony discussion in Section 3.1.3. Notice that even if there is full subscriber participation, so that $n_A + n_B = 1$, whenever $p_i \neq \hat{p}_i$ there are what Laffont, Rey, and Tirole (1998b) term "tariff mediated network externalities" present, and subscribers will choose their network partly on the basis of the number of other

[81] More generally, one can show that whenever there is full subscriber participation, subscriber decisions depend only upon the *difference* in utilities $u_i - u_j$.

[82] The case of fully nonlinear pricing is considered below in Section 4.2.3.

[83] See Kim and Lim (2001), Hermalin and Katz (2001) and Jeon, Laffont, and Tirole (2001) for analyses of the case where networks can charge for incoming calls.

subscribers on the same network. The system of choices in (62) is consistent provided that

$$n_A = s_A(u_A, u_B); \quad n_B = s_B(u_B, u_A).$$

Therefore, given the pair of tariffs offered by the firms, this system of four equations in four unknowns will, at least in the cases we consider, yield the unique equilibrium subscriber numbers. For instance, in the simple Hotelling model (60) one obtains[84]

$$n_A = 1 - n_B = \frac{m_A - \frac{1}{2w}(f_A - f_B)}{m_A + m_B} \tag{63}$$

as the unique equilibrium, where

$$m_i = \frac{1}{2} + \frac{v(\hat{p}_i) - v(p_j)}{2w}.$$

(Here we require that $0 \leqslant n_A \leqslant 1$).

The *net* number of calls from network i to network j, which we denote by z_i, is

$$z_i \equiv n_i n_j (x(\hat{p}_i) - x(\hat{p}_j)). \tag{64}$$

The function z_i represents what might be considered the "net demand for access" by network i. Note that when networks choose the same call charges—so that in particular $\hat{p}_i = \hat{p}_j$—then the net traffic flow between the two networks is zero even if networks have different subscriber numbers.

Turning to the cost side, just as in Sections 3.1 and 4.1 above, let c_i^O be network i's cost of supplying a call which originates on its network and which is terminated on the rival network (not including the termination charge), let c_i^T be network i's cost of terminating a call from the other network, and let the cost of originating and terminating a call entirely within network i just be $c_i^O + c_i^T$. (Thus, there is no cost advantage in carrying a call over a single network rather than over two.) Let k_i be the fixed cost of connecting any subscriber to network i. Therefore, as in Section 3.1, if a subscriber makes q calls and receives Q calls, the total physical costs for network i are $c_i^T Q + c_i^O q + k_i$.

[84] See Section 3 of Laffont, Rey, and Tirole (1998b) for more details, including a discussion of the important issue of whether this equilibrium is stable.

Let a_i be the charge for terminating a call on network i. Therefore, given the two retail tariffs, the resulting market shares, and the two termination charges, total profits for network i are

$$\Pi_i = \underbrace{n_i\{[p_i - c_i^O - c_i^T]n_i x(p_i) + [\hat{p}_i - c_i^O - a_j]n_j x(\hat{p}_i) + f_i - k_i\}}_{\text{profit from subscription}} + \underbrace{n_i n_j (a_i - c_i^T)x(\hat{p}_j)}_{\text{profit from termination}}. \tag{65}$$

Similarly to the function π_i in Section 4.1, introduce the notation

$$\pi_i(p_i) \equiv (p_i - c_i^O - c_i^T)x(p_i)$$

for the profit function for on-net calls. Then we can rearrange (65) to give

$$\Pi_i = n_i\{n_i\pi_i(p_i) + n_j\pi_i(\hat{p}_i) + f_i - k_i\} + n_i n_j\{(a_i - c_i^T)x(\hat{p}_j) - (a_j - c_i^T)x(\hat{p}_i)\}$$

which, when termination charges are reciprocal (i.e. when $a_A = a_B = a$), simplifies to

$$\Pi_i = n_i\{n_i\pi_i(p_i) + n_j\pi_i(\hat{p}_i) + f_i - k_i\} - (a - c_i^T)z_i. \tag{66}$$

As in Section 4.1, the move order is that access charges a_i are chosen first and then, taking these as given, firms choose their retail tariffs non-cooperatively[85].

4.2.2. The first-best outcome

Here we consider the benchmark case where the regulator can control the two firms' retail tariffs directly, and so we calculate the socially optimal two-part tariffs, as in (61), for the two firms. From (65), total industry profits with a given pair of retail tariffs are

$$n_A\{[p_A - c_A^O - c_A^T]n_A x(p_A) + [\hat{p}_A - c_A^O - c_B^T]n_B x(\hat{p}_A) + f_A - k_A\} + n_B\{[p_B - c_B^O - c_B^T]n_B x(p_B) + [\hat{p}_B - c_B^O - c_A^T]n_A x(\hat{p}_B) + f_B - k_B\}.$$

[85] A subtle point is that one has to take care about the choice of strategic variables for the firms when network effects are present. For instance, in (62) above there is a one-for-one relationship between utility u_i and the fixed charge f_i. Given $\{p_i, \hat{p}_i\}$ two competitive scenarios are (i) firms offer utilities u_i and then choose f_i *ex post* in order to deliver the promised utility (which would then depend on the market shares achieved), and (ii) firms offer the fixed charges f_i, and consumers predict the equilibrium market shares and choose their network accordingly. Unfortunately, the outcome is different in the two cases. When this is an issue we will assume that competition takes the form (ii), so that firms compete in tariffs rather than utilities, since that is (perhaps) economically the more plausible.

Using the notation $w(p,c) \equiv x(p)(p-c)+v(p)$ and $c_{ij} = c_i^O + c_j^T$ for the cost a making a call from network i to j, and substituting for f_i in (62), this expression for total industry profits can rewritten as

$$n_A\{n_A w(p_A, c_{AA}) + n_B w(\hat{p}_A, c_{AB}) - u_A - k_A\} + n_B\{n_A w(\hat{p}_B, c_{BA}) + n_B w(p_B, c_{BB}) - u_B - k_B\}.$$

Given the two utilities u_A and u_B offered by the firms, let total subscriber surplus be $V(u_A, u_B)$. This function satisfies the usual envelope conditions

$$\frac{\partial}{\partial u_A} V(u_A, u_B) = s_A(u_A, u_B); \quad \frac{\partial}{\partial u_B} V(u_A, u_B) = s_B(u_A, u_B).$$

Total welfare is therefore

$$W = V(u_A, u_B) + n_A\{n_A w(p_A, c_{AA}) + n_B w(\hat{p}_A, c_{AB}) - u_A - k_A\} + n_B\{n_A w(\hat{p}_B, c_{BA}) + n_B w(p_B, c_{BB}) - u_B - k_B\}.$$

Clearly, for any pair of utilities $\{u_A, u_B\}$, this expression is maximized by choosing each firm's call charges to maximize the relevant welfare function $w(\cdot, c_{ij})$, i.e., to equal the relevant marginal cost:

$$p_i = c_i^O + c_i^T; \quad \hat{p}_i = c_i^O + c_j^T. \tag{67}$$

In particular, when termination costs differ on the two networks, we see that price discrimination according to the destination network is socially optimal.

With the prices in (67) total welfare becomes

$$W = V(u_A, u_B) + n_A\{n_A v(c_{AA}) + n_B v(c_{AB}) - u_A - k_A\} + n_B\{n_A v(c_{BA}) + n_B v(c_{BB}) - u_B - k_B\}.$$

Maximizing the above with respect to u_i, and making the substitution in (62), one obtains the following expressions for the optimal fixed charges[86]:

$$f_A = k_A - n_A v(c_{AA}) - n_B v(c_{BA}); \quad f_A = k_A - n_A v(c_{AB}) - n_B v(c_{BB}). \tag{68}$$

(Of course, the subscriber numbers n_A and n_B are endogenous in the above, and are determined jointly with f_A and f_B).

In sum, (67) shows that call charges should equal the relevant marginal costs, while (68) shows that the fixed charge should be subsidized below the fixed cost to reflect the externality that a subscriber's choice of network exerts on other callers. (If a subscriber chooses to join network A, say, then each of A's other subscribers

[86] Since there is no strategic interaction between the firms in this welfare analysis, the issue mentioned in the previous footnote does not arise, and maximizing welfare with respect to fixed charges and with respect to utilities yields the same result.

obtains benefit $v(c_{AA})$ and each of B's subscribers obtains benefit $v(c_{BA})$, which therefore yields the required subsidy in (68).) In particular, we see that—except in a special case discussed below—it is not optimal to have marginal cost pricing for all services, and the charge for joining a network should differ from cost in order, roughly speaking, to attract more subscribers onto the network with the lower *termination* cost.

The formula (68) is valid for all network choice functions $s_i(u_A, u_B)$, and in general this first-best policy calls for subsidies to be provided out of public funds. However, in the important special case where total subscriber numbers are fixed, so that $n_A + n_B \equiv 1$, then policy is simplified. In particular, since all subscribers opt to join one or other network, market shares depend only on the difference in utilities. Therefore, adding a constant amount to both networks' fixed charge in (68) will not affect subscriber decisions, or total welfare, and so the first best can be achieved without recourse to subsidies.

This special case of full participation also makes clear that it is the difference in the firms' termination, rather than origination or fixed, costs that is the motive for the distortion in (68). For instance, suppose that the firms have the same termination cost, and so in particular that $c_{iA} \equiv c_{iB}$. Then the first-best is achieved by pricing all services, including the fixed charge of joining a network, at marginal cost. In particular, in the symmetric case when the two firms have the same costs—which is the focus of the next section—the optimal outcome is indeed implemented by marginal-cost pricing. The reason for this is that, when call charges are given by (67), differences in origination costs are fully "internalized" by subscribers at the time they choose their network.

By contrast, when firms differ only in their termination costs, then with a marginal-cost pricing regime, subscribers do not fully take into account the implications of their network choice on their callers. (This situation is somewhat related to the model of the mobile market in Section 3.1 above.) To correct for this, the optimal fixed charge/fixed cost margin is lower on the network with the lower termination cost, as in (68).

A natural question to ask is whether this first-best outcome can be implemented by means of a suitable choice of termination charges, and then leaving the two networks to compete at the retail level. Unfortunately, the answer to this, except in a few special cases, is *No*. (In the next section, where the focus is on symmetric situations, we will see that the first-best can often be implemented when the termination charge is equal to cost.) The reason is that access charges are being called upon to perform too many tasks. The first best involves the correct choice of six variables (the two pairs of call charges for the two firms together with the two fixed charges), and yet we have only two instruments (the pair of termination charges) at our disposal. In fact, under a cost-based termination charge regime, where $a_i \equiv c_i^T$, then the four call charges will be at the first-best level, given in (67), in equilibrium. However, there is no reason to believe that the fixed charges that emerge from competition will be equal to (68).

No price discrimination. The above analysis assumed that firms used the most general tariffs as in (61). Since this form of price discrimination involving different call charges for on-net and off-net calls is not common at present, it is worthwhile to perform this welfare analysis under the assumption that firms must charge the same for all calls (i.e., that $p_i \equiv \hat{p}_i$). For simplicity, suppose that there is full subscriber participation, so that $n_A + n_B \equiv 1$. Then one can show that (67) becomes

$$p_A = n_A c_{AA} + n_B c_{AB}; \quad p_B = n_A c_{BA} + n_B c_{BB}.$$

Therefore, the call charge on network i is set equal to the *average* marginal cost as weighted by market shares. And, corresponding to (68), when call charges must be uniform we see that the optimal pattern of fixed charge/fixed cost markups is given by

$$\begin{aligned}(f_A - k_A) - (f_B - k_B) &= n_A x_A [c_{AA} - c_{AB}] + n_B x_B [c_{BA} - c_{BB}] \\ &= [n_A x_A + n_B x_B][c_A^T - c_B^T] \end{aligned} \tag{69}$$

where we have written $x_i = x(n_A c_{iA} + n_B c_{iB})$ for the equilibrium number of calls made by a subscriber on network i.

If the right-hand side of (69) is negative, then network A is subsidized relative to network B. Clearly, this is optimal if the cost of call termination on network A is lower than on B. The intuition for this is precisely the same as for the previous case in (68) where networks offered discriminatory tariffs: with marginal cost pricing, subscribers do not adequately take account of the effect their choice of network has on their callers' call charges. Therefore, incentives to join a network are distorted away from marginal costs in order to induce more people to join the network with the lower termination cost, as shown in (69).

4.2.3. Symmetric competition: A danger of collusion?

Can networks that compete for subscribers be relied upon to choose termination charges that are close to socially optimal? Or can networks use termination charges to affect their equilibrium behaviour in the retail market to boost profits at the expense of welfare? We will see in the following succession of models that the answers to these questions are rather subtle, and depend (i) on whether networks use linear rather than nonlinear tariffs and (ii) on whether they price discriminate between on-net and off-net calls.

In this section we suppose that the two networks are symmetric in terms of cost, so that $c_A^O = c_B^O = c^O$, $c_A^T = c_B^T = c^T$ and $k_A = k_B = k$. Write

$$\pi(p) \equiv x(p)(p - c^O - c^T) \tag{70}$$

for the on-net profit function for each network. We also use the (symmetric) Hotelling formulation for subscriber choices as in (60). Because of symmetry we assume that networks agree to charge a reciprocal access charge a to each other[87].

Linear non-discriminatory pricing. Here, despite its lack of realism, suppose that the networks are constrained to offer only linear tariffs, so that $f_i = 0$. We also suppose that networks are prohibited from engaging in price discrimination according to the destination network, so that $p_i \equiv \hat{p}_i$. In this case, since the total number of subscribers is always equal to 1, (66) simplifies to

$$\Pi_i = n_i\{\pi(p_i) - k\} - (a - c^T)z_i. \tag{71}$$

The joint-profit maximizing (or collusive) linear retail price, denoted p^*, is the price that maximizes π in (70).

Consider the sub-game in which firms choose retail tariffs given that a reciprocal access charge a has initially been chosen. From (71), the first-order condition for p to be the equilibrium choice for each firm is that $\partial\Pi_i(p,p)/\partial p_i = 0$. With the Hotelling specification (60), this entails

$$-\frac{x(p)}{2w}(\pi(p) - k) + \frac{1}{2}\pi'(p) - \frac{1}{4}(a - c^T)x'(p) = 0. \tag{72}$$

Since joint profits are maximized at $p = p^*$, where $\pi'(p^*) = 0$, if collusion *can* be sustained by choosing a suitable access charge, say a^*, from (72) this access charge must be

$$a^* = c^T + \frac{x(p^*)}{-x'(p^*)}\frac{2}{w}(\pi(p^*) - k) > c^T. \tag{73}$$

In particular, we see that $a^* > c^T$. This candidate for the collusive termination charge is high when (i) $\frac{1}{w}$, which is a measure of the substitutability of the two networks' services, is high, (ii) when demand is inelastic, i.e. when $-x/x'$ is high, or (iii) when $\pi(p^*) - k$, the maximum profit per subscriber, is high. Note that in the limit as w becomes very large, so that market shares are fixed, the collusive termination charge is equal to cost. This is an instance of the result in Section 4.1

[87] It is intuitive that if networks set access charges non-cooperatively, they will set them *higher* than if they negotiate over a common charge, since there will be a serious "double marginalization" problem. This is exactly comparable to the case of international call termination discussed in Section 4.1.1 above. See also Laffont, Rey, and Tirole (1998a) for more details. It will therefore be socially desirable to allow firms to "collude" over the choice of the access charge compared to the case where each firm acts independently.

that in symmetric situations the ideal reciprocal termination is equal to cost—see (59) above.

When the access charge is set according to the rule (73), firms have no *local* incentive to deviate from the collusive price p^* in the retail market, even though retail prices are set non-cooperatively: if one firms undercuts the other by a small amount, the gain in retail profits from increased market share is just offset by the increased access payments needed for the increased number of calls made to the rival network.

The remaining question is when the first-order condition (72) does in fact characterize the (globally) optimal response for one firm given that the other has chosen p^*. It is easy to see that services being close substitutes will make this collusion impossible. From (60), if one network deviates and chooses the price $p_L < p^*$, it can take the whole market provided that $v(p_L) \geqslant v(p^*) + w$. Suppose also that this price p_L is fairly close to p^* in the sense that $\pi(p_L) - k > \frac{1}{2}[\pi(p^*) - k]$. If such a price may be found—and clearly this is possible if w is sufficiently small—then the collusive price p^* cannot be sustained in equilibrium with any access charge, for it would always pay a firm to corner the market by choosing the price p_L (which thereby eliminates the need for access payments altogether)[88]. On the other hand, if products are not close substitutes then the sacrifice in retail profits needed to come close to capturing all subscribers is too great and no undercutting of the collusive price p^* is unilaterally profitable.

Therefore, if the services offered by the two networks are sufficiently differentiated, setting a high termination charge may indeed be used as an instrument of collusion. The reason for this is that high termination charges increase the cost of reducing retail prices unilaterally, since such an action causes the deviating network to have a net outflow of calls to the rival network, which then results in costly call termination payments. In sum, a major contrast between the one-way and two-way models of access pricing is that in the former the chief danger is that high access charges can be used as an instrument of *foreclosure*—perhaps to drive out or otherwise disadvantage rivals in the downstream market—whereas in the latter case high access charges can sometimes be used as an instrument of *collusion*[89].

[88] More precisely, if a departs significantly from termination cost then equilibrium in the retail market simply does not exist when the market is sufficiently competitive—see Laffont, Rey, and Tirole (1998a) and Laffont, Rey, Tirole (1998b) for several instances of this kind of problem. An interesting feature of the recent analysis in Laffont, Marcus, Rey, and Tirole (2001), which was discussed in Section 3.2 above, and Jeon, Laffont, and Tirole (2001) is that, even in perfectly competitive markets, equilibrium exists when the access charge departs from cost. In these two latter papers networks charge for receiving calls, whereas in the earlier papers this price was missing. Introducing this price makes competition more "stable"—see the discussion in Section 2 of Laffont, Marcus, Rey, and Tirole (2001) for details.

[89] This kind of analysis can also be applied to the pay-TV market—see Section 2 of Armstrong (1999) for a model in which two TV companies agree to pay each other large amounts for each other's programmes in order to weaken competition for subscribers.

An implication is that firms should not be free to set their termination charges, even in the case where there is no dispute between firms.

In fact this result is not really surprising. It is very plausible that different termination charges will affect the choice of retail tariffs offered by competing networks in equilibrium. For instance, when price discrimination is banned—so that $\hat{p}_i = p_i$—the choice of a directly affects the average cost of a call, and this will naturally feed through into a higher equilibrium call charge. What is more surprising is that this result—that access charges can implement collusion—is easily overturned in natural extensions to this model, as will be seen in the next sections.

Two-part tariffs without network-based price discrimination. Next consider the more realistic case where networks compete using two-part tariffs. (We maintain the assumption that price discrimination according to destination network is banned.) We will show that the collusive impact of the access charge is now eliminated, at least in the simple symmetric and full-participation model we are using. Indeed, it is straightforward to show that whenever a symmetric retail equilibrium exists, the resulting profits cannot depend at all on the choice of termination charge.

With a given reciprocal termination charge a, network i's profit in (71) is modified to be

$$\Pi_i = n_i\{\pi(p_i) + f_i - k\} - (a - c^T)z_i. \tag{74}$$

Suppose that the symmetric retail equilibrium involves each network offering the two-part tariff $T(x) = px + f$, in which case the equilibrium subscriber utility is $u_A = u_B = v(p) - f$. Consider what happens if one firm decides to vary its fixed charge f a little, keeping the call charge constant at p. (Clearly at any equilibrium, a firm should have no incentive to deviate from f.) Because each network offers the same call charge p, the net demand for access z_i in (64) is zero, even when firms have different market shares. Therefore, firms can compete for market share—by reducing f—*without* incurring call termination payments, and this is the major difference with the linear pricing case.

It is now straightforward to derive equilibrium profits. With the market share specification (60), when a firm offers the fixed charge $\hat{f}$ rather than the equilibrium charge f, its total profit in (74) is

$$\left(\frac{1}{2} + \frac{f - \hat{f}}{2w}\right)(\pi(p) + \hat{f} - k).$$

For f to be an equilibrium this expression must be maximized at $\hat{f} = f$, and so we obtain the first-order condition

$$\pi(p) + f - k \equiv w. \tag{75}$$

Since the left-hand side of this expression is the total industry profit (which is split equally between the two networks), we see that the choice of termination charge a has no effect on profits. In particular, a cannot be used as a collusive device when non-discriminatory two-part tariffs are used.

The choice of a *does* affect the equilibrium choice of the price p, however. Indeed, Proposition 7 in Laffont, Rey, and Tirole (1998a) shows (using the current notation) that the equilibrium call charge is

$$p = c^O + \frac{1}{2}(c^T + a), \tag{76}$$

which is just the perceived average cost of a call when networks divide the market equally[90]. Therefore, just as in the above linear pricing case, a high access charge feeds through into a high retail call charge, and hence leads to high profitability from *calls*. However, the additional instrument of the fixed charge is then used to attract these profitable subscribers, with the result that these excess profits are partially competed away[91]. Clearly, in this model the socially optimal retail price is $p = c^O + c^T$, and so (76) implies that the socially optimal access charge is equal to cost: $a = c^T$. Since the firms are indifferent between termination charges when two-part tariffs are used, in theory they will not object to a regulatory suggestion to set $a = c^T$. In this sense, then, competition for subscribers greatly simplifies the regulatory problem when two-part tariff are used.

Two-part tariffs with call externalities. To see how robust the above "profit-neutrality" result is, consider the following extension. Suppose that each subscriber obtains utility b for each call received. As above, networks charge a two-part tariff without price discrimination according to destination network. In this case subscriber utility in (62) is modified to

$$u_i = v(p_i) - f_i + b[n_i x(p_i) + n_j x(p_j)] .$$

Here, the term in $[\cdot]$ is the total number of calls received by a subscriber, and this is *independent* of the chosen network. Thus, a subscriber's utility from receiving calls, which does in general depend on the market shares of the two networks, does not bias a subscriber's network choice at all[92]. Since with the Hotelling

[90] This proposition also shows that equilibrium exists when w is not too small or $a - c^T$ is not too large.

[91] This effect is similar to that seen in the analysis of the mobile market in Section 3.1. There, networks make large profits from terminating calls which are then (fully) competed away at the retail level.

[92] In the analysis by Jeon, Laffont, and Tirole (2001) of call externalities and the effects of charging for incoming calls, this feature of the market plays a central role. Call externalities represent a *direct* externality, whereas charging for incoming calls creates a *pecuniary* externality. In most cases, those authors show that only the latter effect is relevant for firms' and subscribers' incentives.

specification in (60) market shares depend only on the difference in offered utility levels, the presence of call externalities has no effect on the competitive outcome. In particular, the equilibrium profits and retail tariffs are still given by (75) and (76), and are not affected by the externality parameter b.

What is affected, however, is the socially optimal termination charge. With the call externality b, the socially optimal call charge is reduced to $p = c^O + c^T - b$ in order to stimulate call volume. From (76) this implies that the socially optimal termination charge is below cost:[93]

$$a = c^T - 2b \ .$$

Again, though, the fact that profits are unaffected by the termination charge suggests that firms should not object to this regulatory requirement to subsidize network interconnection[94].

Nonlinear pricing with heterogeneous subscribers:[95] Another important extension to the basic analysis is to introduce more differences in subscribers. The analysis until now has assumed homogeneity in the calling patterns of subscribers. In particular (i) all subscribers made the same number of calls and (ii) all subscribers received the same number of calls. In this section we extend the analysis to allow for subscriber heterogeneity. Again, though, we do not permit networks to discriminate according to the destination network. For simplicity, we return to the case where subscribers do not care about receiving calls (i.e. $b = 0$).

Specifically, subscribers can differ in their demand for calls, and a *high* (respectively *low*) demand type of subscriber is denoted by H (L). The fraction of subscribers with high demand is α. It is possible that high demand subscribers also differ in the number of calls they receive, and suppose that a fraction β^k of the calls made by type k subscribers are made to type H subscribers, and the fraction $1 - \beta^k$ of their calls are made to low demand consumers. (If $\beta^k \equiv \alpha$ then we would have a model where subscribers were equally likely to call each other subscriber, regardless of the demand characteristics of the call recipient.) Because of this heterogeneity, networks will offer a pair of contracts, one for each of the two kinds of subscriber. Without loss of generality, these contracts specify a total

[93] Comparing this reduction to that in the mobile sector—see expression (46)—we see that the subsidy is required to be twice as great. The reason for this is that, in the current context, it is required that a network's perceived (average) marginal cost for a call be reduced by b. However, since only half of a network's calls are destined for the rival network—and so only half incur the cost a—the reduction in a must be correspondingly amplified.

[94] When the analysis is extended to allow for charging for incoming calls, as in Jeon, Laffont, and Tirole (2001), then there are two available instruments for implementing the desired call charge. For instance, one could have call termination at cost, and impose a suitable incoming call charge. When, however, there is the possibility that some subscribers will refuse to accept incoming calls rather than pay the incoming call charge, then it is superior not to charge for incoming calls.

[95] This is adapted from Dessein (2000) and Hahn (2000). See also Section 3.5 of Tirole (1988) for an account of the approach to nonlinear pricing used here.

number of calls X in return for a charge f. Suppose that network i offers the type k subscriber a contract allowing X_i^k calls in return for a charge f_i^k. We assume that in equilibrium both demand types of subscriber are served, so that we do not have to consider network externality effects[96].

We suppose that network choices continue to be made according to the Hotelling model, i.e. that if u_i^k is the *maximum* utility that a type k subscriber can obtain from the contracts offered by network i, then the fraction of type k subscribers who subscribe to network i is given in (60). (Implicitly, we are assuming that the "vertical" differentiation parameter $k = L, H$ is uncorrelated with the "horizontal", or brand preference, parameter y. Therefore, knowledge of brand preference tells a firm nothing about a subscriber's likely demand for calls.) It actually makes no difference to the following argument whether or not networks can observe the demand type of any given subscriber[97]. However, if a subscriber's demand characteristics *are* private information then networks must ensure that the incentive constraints are satisfied, and that a type k subscriber finds it privately optimal to choose the contract (X_i^k, f_i^k) instead of the contract aimed at the other demand type.

If network i has a fraction n_i^k of the demand type k subscribers, the total number of calls received by a type H subscriber (on any network) from callers on network j can be shown to be

$$n_j^H \beta^H X_j^H + \frac{1-\alpha}{\alpha} n_j^L \beta^L X_j^L \ ,$$

while the number received by a type L subscriber (on any network) from callers on network j is

$$\frac{\alpha}{1-\alpha} n_j^H (1-\beta^H) X_j^H + n_j^L (1-\beta^L) X_j^L \ .$$

The number of calls made to network j by a type k subscriber on network i is

$$X_i^k [n_j^H \beta^k + n_j^L (1-\beta^k)] \ .$$

Therefore, with these call volumes and market shares, network i's net outflow of calls—i.e. its demand for access z_i—is given by the complex expression

$$\begin{aligned} z_i =& \alpha n_i^H X_i^H [n_j^H \beta^H + n_j^L (1-\beta^H)] + (1-\alpha) n_i^L X_i^L [n_j^H \beta^L + n_j^L (1-\beta^L)] \\ & - \alpha n_i^H \left[n_j^H \beta^H X_j^H + \frac{1-\alpha}{\alpha} n_j^L \beta^L X_j^L \right] \end{aligned}$$

[96] See Poletti and Wright (2000) for analysis of the case where participation constraints have an impact on networks' policies. They show that the profit neutrality result fails in this case.

[97] Or, in the terminology often used in the literature, it makes no difference whether the model is one of third degree or second degree price discrimination.

$$-(1-\alpha)n_i^L\left[\frac{\alpha}{1-\alpha}n_j^H(1-\beta^H)X_j^H+n_j^L(1-\beta^L)X_j^L\right]. \tag{77}$$

If the reciprocal call termination charge is a, then, similarly to (66), network i's profit is

$$\Pi_i=\alpha n_i^H[f_i^H-k-X_i^H(c^O+c^T)]+(1-\alpha)n_i^L[f_i^L-k-X_i^L(c^O+c^T)]-(a-c^T)z_i$$

where z_i is as in (77).

Notice that when $X_A^k=X_B^k=X^k$ say, so that the two networks offer the same pair of quantities in their contracts, the expression for z_i in (77) simplifies to

$$z_i=(n_i^Hn_j^L-n_i^Ln_j^H)[\alpha X^H(1-\beta^H)-(1-\alpha)X^L\beta^L]. \tag{78}$$

In particular, when we also have $n_i^L=n_i^H$, so that network i does not attract a disproportionate number of one demand type of subscriber, then $z_i=0$.

Next consider equilibrium profits for a given reciprocal access charge a. Given the symmetry between networks, suppose that each network offers the same pair of contracts $\{(X^L,f^L),(X^H,f^H)\}$ in equilibrium. Suppose that network i deviates from this candidate equilibrium by adding an amount ε to both fixed charges f^L and f^H. (In the case where the subscriber demand type k is private information, this uniform change to the fixed charges has no effect on the relative merits of the two contracts, and so does not affect incentive compatibility.) Since we assume that competition causes the participation constraints not to bind for subscribers, this modification does not drive any subscribers from the market, although it obviously drives marginal subscribers from one network to the other. From (60) this modification means that the network loses an *equal* proportion of both types of subscriber[98].

In particular, from (78) we see that $z_i=0$ and the deviating firm incurs no access deficit. Similarly to (63), then, profits with this deviation are

$$\begin{aligned}\Pi_i=&\alpha\left(\frac{1}{2}-\frac{\varepsilon}{2w}\right)\left[f^H+\varepsilon-k-X^H(c^O+c^T)\right]\\&+(1-\alpha)\left(\frac{1}{2}-\frac{\varepsilon}{2w}\right)\left[f^L+\varepsilon-k-X^L(c^O+c^T)\right].\end{aligned}$$

For the pair of contracts $\{(X^L,f^L),(X^H,f^H)\}$ to be an equilibrium, it is necessary that this expression be maximized at $\varepsilon=0$, which, just as in (75), yields the first-order condition

[98] Dessein (2000) looks at the case where different subscriber types also have different transport cost parameters w. For instance, high demand subscribers might be more price sensitive. In this case the argument fails, and the access charge has an effect on profits.

$$\alpha\left[f^H - k - X^H(c^O + c^T)\right] + (1-\alpha)\left[f^L - k - X^L(c^O + c^T)\right] = w\ .$$

The left-hand side of this expression is just the total profits in the market, which is equal to the product differentiation parameter w. Therefore, equilibrium profits again are unaffected by the level of the termination charge a. Thus we see that the profit-neutrality result is not an artifact of the assumption that subscribers were homogeneous.

As in the case of two-part tariffs, the choice of a *does* affect the choice of contracts offered, and $\{(X^H,f^H),(X^L,f^L)\}$ will be a complicated function of the termination charge. The socially optimal choice for a will therefore be the choice that implements the best pattern of consumption. In the case where $a = c^T$, it is possible to show that firms in equilibrium will simply offer the cost-based two-part tariff $T(x) = (c^O + c^T)x + w + k$, so that calls are charged at (actual and perceived) cost, and the firms make a profit of w per subscriber (just as in the previous section on two-part tariffs)[99]. Therefore, as before the socially optimal access charge is $a = c^T$, and since firms are indifferent between all levels of the access charge, in principle they will not object to this regulatory policy.

This analysis of two-part tariffs and nonlinear pricing seems to suggest that the choice of termination charge cannot affect profits at all, and networks will not object to a regulatory suggestion to price interconnection at cost (or below cost in the case of call externalities). However, it is important to stress that this convenient result is non-robust in a number of dimensions. For instance the assumed cost and demand symmetry across networks plays an important role in the argument. (Without this it is unlikely that networks will choose reciprocal access charges.) Another reason for non-neutrality is explored in the next section[100].

Two-part tariffs with network-based price discrimination. Here we return to the homogeneous subscriber framework, but now allow networks to make their call charges depend on the destination network, i.e. to set $p_i \neq \hat{p}_i$. Although this problem looks rather complex, the analysis is simplified by the following observation: in equilibrium all call charges are equal to perceived marginal costs, so that

$$p_A = p_B = c^O + c^T\ ; \quad \hat{p}_A = \hat{p}_B = c^O + a\ . \tag{79}$$

(The reasoning for this is the same as that used to derive (67).)

[99] See Dessein (2000) and Armstrong and Vickers (2001) for further details.

[100] A further framework in which the profit-neutrality result is unlikely to hold is when there is only partial subscriber participation in the market, i.e. when the total number of subscribers rises as the available consumer surplus increases. (This feature was discussed in Section 3.1.3 on the mobile sector.) In this case there will be a "market expansion" effect as well as a "business stealing" effect of the choice of access charge, and the former will most likely make equilibrium profits depend on the access charge. The analysis of this case appears to be somewhat technical, and for some preliminary results see Dessein (2000).

Armed with this observation, expressions (63) and (65) simplify to

$$n_i = \frac{1}{2} + \frac{1}{2}\frac{f_j - f_i}{w + v(c^O + a) - v(c^O + c^T)}$$

and

$$\Pi_i = n_i(f_i - k) + n_i n_j(a - c^T)x(c^O + a) .$$

Therefore, the symmetric equilibrium choice of $f_A = f_B = f$ is given by $f = k + w + v(c^O + a) - v(c^O + c^T)$[101]. The resulting total industry profits are then

$$\Pi = w + v(c^O + a) - v(c^O + c^T) + \frac{1}{2}(a - c^T)x(c^O + a) . \tag{80}$$

Clearly, this profit does now depend on a, and so the profit-neutrality result does not hold when this form of price discrimination is permitted.

When $a = c^T$ we obtain the same profit w as in (75) for the no-discrimination case analyzed above. (In this case firms do not choose to practice price discrimination, even if they are permitted to do so.) However, networks can do better than this when price discrimination is allowed. For profit in (80) is maximized by choosing the termination charge

$$a = c^T - \frac{x}{-x'} < c^T ,$$

and so profits are at their maximum if call termination is *subsidized*[102].

Clearly, social welfare is maximized in this model by setting $a = c^T$, and so there is once again a role for network access regulation in this market[103]. Thus, in direct contrast to the linear pricing case discussed above, where high access charges were used to bolster profits at the expense of welfare, here *low* access charges are privately desirable but socially costly. The intuition for this perhaps surprising result is that when $a < c^T$ it is cheaper for subscribers to call people on the rival network than people on their own network, and so, all else equal,

[101] We are ignoring the details about when such an equilibrium exists—see Proposition 5 in Laffont, Rey, and Tirole (1998b) for a discussion of this.

[102] There is an error in the proof of Proposition 5 in Laffont, Rey, and Tirole (1998b) which led those authors to conclude that choosing $a = c^T$ maximized profits. This has been corrected in Gans and King (2001). The latter paper argues that the "bill and keep" system for network interconnection—where each network delivers the other's traffic for free—might be a reasonable approximation to this profit-maximizing termination charge (especially if traffic monitoring costs are taken into account).

The analysis presented here assumes that firms used the fixed charge f_i as the strategic variable—see footnote 84 above. Intriguingly, if firms instead used utilities u_i as their strategic variable, then one can show that setting $a = c_T$ becomes the profit-maximizing choice.

[103] Alternatively, public policy could prohibit this kind of price discrimination, which acts to restore the profit neutrality result, thereby making the regulatory task easier.

subscribers prefer to belong to the *smaller* network. (Recall that, regardless of market shares, each network sets the same pair of call charges given by (79).) This means that the market exhibits "negative network externalities", and firms have little incentive to compete aggressively for subscribers. In effect, by means of a suitable choice of termination charge firms can coordinate on whether to have a market with positive network externalities (high a), no network externalities ($a = c^T$) or negative network externalities (low a). Because of its softening effect on competition, it is mutually profitable for firms to choose the third option[104].

Discussion. We have seen that the relationship between access charges and equilibrium profits—and in particular whether access charges can be used to sustain "collusion" —depends, in part, on the kinds of tariffs the firms are able to offer. One way to get an intuition about this is to use the framework of "instruments and objectives" in Section 2.2.3: when do the firms have enough instruments to attain a given (perhaps collusive) outcome? Since our analysis has assumed the use of only a single instrument—the reciprocal access charge a—firms are likely to be able to sustain a given retail equilibrium only in the case of linear pricing. With linear pricing there is only one objective—the maximization of equilibrium profits within the (restrictive) class of outcomes possible with linear tariffs—and this single instrument will often be enough to implement the optimum. (However, we saw that sometimes equilibrium did not exist for a wide enough range of access charges for this argument to work.)

Other kinds of tariffs, such as two-part tariffs, have at least two dimensions to their definition, and so the single instrument of the access charge will not in general be sufficient to induce a given retail equilibrium. In particular, the joint-profit-maximizing two-part tariff will not in general be sustainable by a particular choice for a[105]. However, it is one thing to note that the access charge will not be able to sustain the *maximum* profit, and quite another to obtain the result that the access charge has *no effect* on equilibrium profits. This striking profit neutrality result was obtained for the case of both two-part tariffs, both with and without

[104] Laffont, Rey, and Tirole (1998b) also discuss the case of *linear* pricing with price discrimination. Proposition 2 in that paper shows that when price discrimination is allowed and w is quite large, networks will choose to set the termination charge above cost. Also, Proposition 3 shows that when w is large allowing price discrimination is good for welfare (keeping the termination charge fixed).

[105] If firms did have another instrument at their disposal, then, at least in some circumstances, we would expect collusion to be possible when two-part tariffs are used. For instance, if firms signed a reciprocal agreement so that, say, firm A paid firm B a specified amount for each subscriber A served, then this would be a means of controlling the fixed charge element of the equilibrium two-part tariff. (The access charge then is left to control the usage charge in the usual way.) In a sense, this additional instrument performs a similar role to the "output tax" instrument discussed in Section 2. However, there are a number of reasons—ranging from difficulties involved in monitoring subscriber numbers, to competition law—why such contracts may not be possible.

Another possible instrument is a charge for incoming calls. When this is used, Jeon, Laffont, and Tirole (2001) show that it is sometimes possible for networks to obtain the joint profit-maximizing outcome with a suitable choice of termination charge and reception charge.

call externalities, and for fully nonlinear pricing. (However, it did not hold for the case of network-based price discrimination). Although there has been little work done, so far, outside the frameworks discussed in this section, one might conjecture that this neutrality result is *very* special, and will depend, for instance, on the specific Hotelling subscriber choice rule we used (involving full subscriber participation). If so, then there will remain a role for regulation of the termination charge.

4.2.4. Asymmetric competition and non-reciprocal access charges

The previous section analyzed symmetric competition between networks, and so might be relevant to situations where competition is well-established and mature. In earlier stages of market liberalization, however, competition if left to itself is likely to be skewed in favour of the incumbent, and the analysis needs to be extended to cover such, often more relevant, situations. Unfortunately, there have been only few contributions to the theory of asymmetric competition between networks, so the following discussion is more speculative and incomplete.

Perhaps the central reason why a proper analysis of asymmetric markets is so hard is that there is no reason *a priori* to suppose that access charges should—either from the firms' point of view or in terms of overall welfare—be set reciprocally[106]. Even if the two networks' termination costs (or other costs) were assumed to be the same—itself a questionable assumption if one firm is established and the other is an entrant—this assumption of reciprocity is not innocuous. For instance, in the model with a regulated incumbent firm described below the regulator does not want generally to choose termination charges reciprocally, even when termination costs are the same for the two networks. The discussion of international call termination in Section 4.1.2 was also in the context of asymmetric networks, but that framework is a good deal more straightforward than this case where firms compete for subscribers. However, even in that simple framework we saw that the outcome depended on the details of the bargaining process generating the equilibrium. The advantage of symmetry in Section 4.2.3 is that the firms' interests coincide at the stage where access charges are determined, and so this bargaining stage is trivial. In asymmetric situations the details of the procedure for choosing access charges will be more important. One of the most valuable areas for future research will be to find compelling theoretical models of how non-reciprocal access charges are determined: ideally tractable models can be found; if not, then numerical simulations seem likely to be most promising way forward[107].

[106] Another reason why the asymmetric analysis is not transparent is that marginal cost pricing is then generally not the first-best outcome—see Section 4.2.2 above.
[107] Numerical simulations are used extensively in de Bijl and Peitz (2002).

In the remains of this section we briefly discuss two related models[108]. The first is a model where the two networks choose their retail tariffs without regulatory control, and the second is a model where the dominant firm's retail tariff is controlled.

A simple model of asymmetric competition and non-reciprocal termination charges. The crucial simplification that we make here, and in the next model, in order to make the analysis tractable is to suppose that subscribers have *inelastic* demand for calls[109]. Specifically, suppose that, regardless of the price per call, each subscriber wishes to make exactly one unit of calls to each other subscriber. (We also assume there is full subscriber participation, so that $n_A + n_B \equiv 1$.) In this case, all that matters for a subscriber's network decision is the total charge a network levies for making a single unit of calls to all other subscribers[110]. Suppose, then, that this combined charge is P_i on network i. If $\bar{u}$ is some (high) gross utility a subscriber receives from making his calls, then the utility received by a subscriber if she joins network $i = A, B$ in (62) becomes

$$u_i = \bar{u} - P_i \,. \tag{81}$$

Let a_i be the charge for terminating a call on network i. Therefore, total profits for network i are simplified from (65) to become

$$\Pi_i = \underbrace{n_i\{P_i - [c_i^O + c_i^T]n_i - [c_i^O + a_j]n_j - k_i\}}_{\text{profit from subscription}} + \underbrace{n_i n_j (a_i - c_i^T)}_{\text{profit from termination}} \tag{82}$$

Introducing the notation

$$C_i \equiv c_i^O + c_i^T + k_i \,,$$

[108] Other treatments of asymmetric competition are Section 7 in Laffont, Rey, and Tirole (1998a) and Section 6 in Laffont, Rey, Tirole (1998b), although the focus there is mainly on the case of reciprocal access pricing. The source of the asymmetry there is the fact that the incumbent has full geographic coverage, whereas the entrant has to install coverage (with a convex cost in so doing). Carter and Wright (2001) present a model of vertical differentiation, in which the incumbent offers a superior service to the entrant. (The cost functions of the two firms do not differ.) Again, though, the analysis focuses on reciprocal access charges. They find that the "larger" firm would like the reciprocal access charge to be equal to the firms' termination cost. Moreover, they show that when the market is very asymmetric, the smaller firm would also like the access charge to be set equal to cost.

[109] This section has particularly benefitted from comments from Julian Wright. This section is closely related to Section 6.3 of de Bijl and Peitz (2002).

[110] In fact, this is not quite true. If firms charged differently for on-net and off-net calls, then subscribers would have to come to a view about the equilibrium market shares before they can choose their network. Similarly to the discussion in footnote 85 above, this would affect the equilibrium. For simplicity, we side-step this issue and assume that firms compete in 'total charges' P_i which are invariant to the realized market shares.

and using the assumption that $n_A + n_B \equiv 1$, implies that we can rearrange (82) to give

$$\Pi_i = n_i\{P_i - C_i\} + n_i n_j\{a_i - a_j\} \ . \tag{83}$$

In particular, the outcomes (prices, profits, market shares, welfare) depend only on the *difference* between termination charges on the two networks.

In order to introduce a demand-side asymmetry between the two networks, we can modify the symmetric Hotelling formulation in (60) to give a bias in favour of, say, firm A[111]. Therefore, suppose that if the two firms' utilities are u_A and u_B, the subscriber located at y obtains utility $u_A - ty + t\beta$ if she joins network A, and utility $u_B - t(1-y)$ if she joins network B. If $\beta > 0$ then, with equal utilities offered, firm A will obtain the higher market share. With the two utilities $u_i = \bar{u} - P_i$, the market shares n_i are then given by

$$n_A = \frac{1+\beta}{2} + \frac{P_B - P_A}{2w} \ ; \ n_B = \frac{1-\beta}{2} + \frac{P_A - P_B}{2w} \ . \tag{84}$$

After substituting these market shares into (83), given an initial choice of termination charges $\{a_A, a_B\}$ some tedious calculations show that the equilibrium prices for the two firms are

$$\begin{aligned} P_A &= \frac{2C_A + C_B + \beta w}{3} + w + \frac{\Delta^a}{3}\left\{\beta - \frac{\Delta^C}{w}\right\} \\ P_B &= \frac{C_A + 2C_B + \beta w}{3} + w + \frac{\Delta^a}{3}\left\{\beta - \frac{\Delta^C}{w}\right\} \end{aligned} \ . \tag{85}$$

Here, we have written $\Delta^C = C_A - C_B$ for the cost difference and $\Delta^a = a_A - a_B$ for the difference in termination charges.

Notice that the termination charge differential, Δ^a, affects the two firms' equilibrium prices in the *same* way. This implies that the equilibrium price difference, $P_A - P_B$, is

$$P_A - P_B = \frac{1}{3}\{\Delta^C + 2\beta w\} \ . \tag{86}$$

Crucially, this price difference, which is what determines the equilibrium market shares, does *not* depend on the termination charges. The equilibrium market shares of the two firms are

[111] This is the specification for network choice used in Carter and Wright (1999) and Carter and Wright (2001). The following analysis does not depend at all on this particular Hotelling specification, and will apply to any case where there is full participation. We use this functional form merely to derive a closed-form equilibrium.

$$n_A = \frac{1}{2} + \frac{1}{6}\left\{\beta - \frac{\Delta^C}{w}\right\} \; ; \; n_B = \frac{1}{2} - \frac{1}{6}\left\{\beta - \frac{\Delta^C}{w}\right\} . \tag{87}$$

Without loss of generality, suppose we label the firms so that

$$\beta w > \Delta^C . \tag{88}$$

This implies that A's advantage on the demand side—all else equal, subscribers are willing to pay βw more for A's service than for B's—outweighs any cost disadvantage Δ^C. From (87) this implies that A has the larger market share in equilibrium. In addition, for the preceding analysis to be valid we require that the larger firm does not 'corner the market', i.e., that $n_A < 1$ at the equilibrium. Clearly, this requires that the asymmetries are not too great, in the sense that parameters satisfy

$$\beta w - \Delta^C < 3 . \tag{89}$$

When (88) holds, (85) implies that *both* retail prices increase as firm A is paid more than B for call termination. In this sense, non-reciprocal access charges, with the larger firm being paid a higher access charge, act as an 'instrument of collusion', in a similar way to the linear pricing model in Section 4.2.3. The reason why having a large difference Δ^a acts to relax competition for subscribers is quite intuitive. From (83), when Δ^a is large firm A would like to *increase* the volume of off-net traffic. However, off-net traffic, which is just $n_A n_B$ in this model, is maximized when market shares are equal, and so this gives the larger firm an incentive to make the market more symmetric, i.e., to compete less hard. On the other hard, the smaller firm would like to *decrease* the volume of cross-network traffic as it makes a loss on this traffic when Δ^a is large. This implies that the smaller firm would like to move further from the symmetric allocation, which again gives it an incentive to compete less hard. This is why an increase in Δ^a induces both firms to raise their prices.

However, and in contrast to the symmetric analysis of Section 4.2.3, high retail prices are not sufficient to ensure that both firms' profits are high, since we have to take into account the payments for call termination as well. (With non-reciprocal charges, payments for network access do not cancel out in equilibrium, even if off-net traffic is the same in the two directions.) Since termination charges do not affect market shares, it is straightforward to derive the effect on profits for each firm. Since P_A increases with Δ^a, and also, from (83), the profits of the larger firm A are directly increasing in Δ^a, it follows that this larger firm would like the difference Δ^a to be set as large as possible. However, although the retail price of the smaller firm B also increases with Δ^a, from (83) the direct effect of increasing Δ^a is negative. To derive the net effect of increasing Δ^a on B's equilibrium profits, note that its profits decrease with Δ^a whenever $P_B - n_A \Delta^a$ decreases with Δ^a

(where P_B is an increasing function of Δ^a). However, this is the case whenever (89) holds, which we have assumed. Therefore, the smaller firm would like the difference Δ^a to be set as *low* as possible.

In sum, as one might expect, the two firms have divergent preferences over how termination charges should be set, and regulation of some form is likely to be needed to resolve disputes. (However, since total industry profits do increase with Δ^a, it would be profitable for the larger firm to compensate by means of a side payment the smaller firm in order to induce the latter to accept a high Δ^a.)

Although total welfare (profits plus subscriber utility) is affected by the market shares of the two networks, for a *given* market share the inelastic demand assumption means that welfare is constant. Therefore, since termination charges cannot be used to alter market shares in this model, total welfare also is unaffected by the choice of termination charges. Thus, in contrast to the symmetric situations analyzed in the previous section where a profit-neutrality result was derived, in this asymmetric model we obtain a *welfare*-neutrality result.

However, consumer welfare in isolation is affected by the choice of termination charges, and both retail prices are an increasing function of Δ^a. Therefore, consumer welfare is increased by choosing a lower value for Δ^a. To illustrate this point, suppose that firm A's advantage stems at least in part from the cost side, so that $C_A < C_B$. In this case, consumer welfare is higher with a "cost-based" termination charge regime—so that $a_A < a_B$—than with a reciprocal charging regime ($a_A = a_B$).

A final question to ask is: how does the equilibrium outcome compare with the first best? Using the arguments used in more general settings in Section 4.2.2, the first best is achieved when[112]

$$P_A - C_A = P_B - C_B \,. \tag{90}$$

However, (86) implies that

$$P_A - C_A = P_B - C_B + \frac{2}{3}\{\beta w - \Delta^C\} \,.$$

Therefore, since we have assumed that firms are labelled so that the term $\{\cdot\}$ is positive, we see that competitive outcome involves the larger firm setting too *high* a price compared to its rival than is socially optimal. Put another way, competition results in an outcome that is not "asymmetric enough" from the point of view of overall welfare.

A model with a regulated incumbent. To provide a second example of asymmetric competition between networks, consider the following model with a regulated

[112] In fact, this is an instance of the second-best pricing rule in (5) above.

incumbent facing a fringe of entrants. (Here we have a fringe of entrants, rather than a single rival, in order to abstract from the issue of the market power of the entrant—see footnote 7 above.) As in Section 2, the fringe is modeled as consisting of a large number of small networks, each offering exactly the same service and having the same cost functions as each other, and where this service is differentiated from that of the incumbent. In all other respects, including the critical assumption of inelastic demand for calls, the model is as presented above. The incumbent firm A is assumed to be regulated at the retail level, and is required to offer the combined charge P_A to its subscribers. This tariff is assumed to be held constant in the following analysis. Its rivals are a competitive fringe, denoted B. This fringe is unregulated at the retail level, and each firm in the fringe offers the combined charge P_B. As before, the utility of a subscriber on the incumbent's network is $u_A = \bar{u} - P_A$, and the utility of those who go to the fringe is $u_B = \bar{u} - P_B$. The market share of the incumbent is n_A, while the fringe takes the remaining $n_B = 1 - n_A$ subscribers.

We look for an access pricing regime that implements the socially optimal outcome, which we know requires that the fringe price satisfies the property in (90). Suppose that the termination charges on A and B are a_A and a_B, respectively. Suppose that all firms within the fringe are "small" in the sense that a negligible fraction of calls that originate on a fringe network is terminated on the same network. In that case, analogously to (82), the profit per subscriber for the fringe is

$$\Pi_B = \underbrace{P_B - \left[c_B^O + a_A\right]n_A - \left[c_B^O + a_B\right]n_B - k_B}_{\text{profit from subscription}} + \underbrace{a_B - c_B^T}_{\text{profit from termination}}$$

which can be re-written as

$$\Pi_B = P_B - C_B - \Delta^a n_A \ .$$

Competition within the fringe implies that each firm makes zero profits. This implies that the equilibrium retail charge offered by the fringe is

$$P_B = C_B + \Delta^a n_A \ .$$

Again, this has the same feature as (85) that B's equilibrium price is an increasing function of Δ^a. However, the reason for this in this case is different: the more a fringe firm receives for terminating calls on its network, the lower its retail charge has to be in order to break even. Therefore, the reason is the same as in the model of the mobile market in Section 3.1, where high termination payments were used to fund retail subsidies in equilibrium.

Since, from (90), we wish to implement the fringe price $P_B = C_B + [P_A - C_A]$, we see that the optimal pair of termination charges satisfies

$$a_A - a_B = \frac{P_A - C_A}{n_A} .$$

In particular, a reciprocal termination charge is optimal when $P_A = C_A$, i.e., when the incumbent is *optimally* regulated. Perhaps counter-intuitively, then, when there are no regulated distortions at the retail level, it is *not* the case that a cost-based termination regime, where $a_i = c_i^T$, is likely to be optimal.

In other cases, however, non-reciprocal termination charges are used to implement the optimal second best price P_B in (90) given the distorted price charged by the incumbent. If the incumbent is profitable in this market ($P_A > C_A$) then it should charge more for call termination than the fringe, regardless of the underlying relative costs of call termination.

5. Conclusion: Instruments and objectives

This chapter has had a number of objectives. In particular, relative to most of the existing writing on access pricing, I have aimed to do the following:

1. Pay more attention to the issue of network bypass. When bypass is taken into account, access charges that differ from the incumbent's costs (for instance, ECPR-style access pricing) might have the unfortunate effect of inducing inefficient use (or lack of use) of the incumbent's network.
2. Relatedly, make a more forceful case for pricing access at cost, with the important proviso that suitable retail-level instruments are made available to correct for the incumbent's retail tariff distortions. (Output taxes or subsidies levied on entrants, perhaps in the form of a universal service fund, were suggested for this purpose.)
3. Relate the contentious ECPR policy more clearly to familiar, and well-accepted, principles concerning the theory of the second best.
4. Observe that the more tasks that the access charge is required to perform unassisted—and in the chapter these included (a) the need to give entrants good make-or-buy incentives, (b) the need to induce a desirable amount of entry, given the incumbent's retail tariff distortions, and (c) the need to control the incumbent's retail prices when these were not directly controlled by regulation—the more complicated, and informationally-demanding, the various access pricing formulas become.
5. Present a unified treatment of the two way access pricing problem, encompassing a number of recent theoretical contributions. In particular, it was clearly seen that "easy" policy conclusions—such as (i) access charges can be used to sustain collusive outcomes or (ii) access charges have no effect on equilibrium profits—were not robust to small changes in the assumptions.
6. Argue that, even in the most "competitive" markets—such as the mobile sector—there is likely to remain a role for regulation of access charges.

The instruments used to try to achieve these objectives have been kept as simple and streamlined as possible. Most importantly, I have discussed only the *theory* of access pricing; difficulties involved in implementation have been conveniently ignored. There are numerous further modelling assumptions that have been made, I believe, merely for presentational clarity and tractability. These include:

Full information about costs. I assumed throughout that the incumbent's costs were known to all, and also could not be affected by, say, managerial cost-reducing effort. While it is clear that imperfect regulatory knowledge of costs and the potential for cost reduction has an important impact on regulatory policy, the *interaction* of these features with the access pricing problem does not often seem to generate many new insights. Thus, it could be said that, to the extent they depend on the realized costs of the incumbent, access pricing regimes such as cost-based access, ECPR or Ramsey pricing all exhibit features of "cost-plus" regulation. In particular, if an incumbent is required to pass on efficiency gains to entrants in the form of lower access charges, it may have poor incentives for cost-reduction. (This is particularly true in the access pricing context, where the incumbent is required to pass on reductions to its *rivals*, rather than merely final consumers as in most monopoly models of optimal regulation.) While this is obviously true, these problems can be tackled using familiar (but imperfect) methods, such as basing access charges on estimated efficient costs, perhaps including computer generated "engineering models" or benchmarking from observed costs in other countries. The global price-cap proposal of Laffont and Tirole is another response to this issue, in that the permitted set of access charges and retail prices do not depend—in the short run, at least—on the firm's realized costs[113].

Constant marginal costs. Similarly, marginal costs were assumed to be unaffected by the scale of output. This makes little difference to the analysis. For instance, with the competitive fringe model in Section 2.3.3. the analysis would go through if the terms C_1 and C_2 were just interpreted as the (endogenous) marginal costs evaluated at the equilibrium[114].

Other assumptions are far from innocuous, however, and the strong focus on *static* analysis is perhaps the leading limitation of the analysis. For instance, all costs of the incumbent were taken to be avoidable if that firm ceased supplying the relevant service. One could take the convenient view that costs such as C_1 and C_2 are just taken to represent the forward-looking, avoidable component of the firm's costs, and as such are the relevant costs when discussing the efficiency of

[113] See Section 4.7 in Laffont and Tirole (2000) for further discussion. See also Laffont and Tirole (1994) for a full account of optimal access pricing with asymmetric information, and DeFraja (1999) for related analysis.

[114] See Armstrong, Doyle, and Vickers (1996) for analysis along these lines.

future entry given that important parts of the incumbent's network are already sunk. This, however, is not at all satisfactory. Were the initial investments made with full knowledge of the future access regime? (If so, then all costs are avoidable *ex ante* and the analysis of this chapter is relevant.) This is unlikely, though, given the often long-lived nature of many infrastructure investments in the industry, together with the typical long-run unpredictabilty of regulatory policy. However, to use *ex post* avoidable costs as basis for access pricing policy (to be determined once investments are sunk) is a recipe for opportunism and "deregulatory takings" of the kind emphasized in the writings of Sidak and Spulber. An important next step for theoretical research in this area is, I believe, to provide a proper analysis of the *dynamics* of access pricing, focussing on the need to provide long-run, stable incentives for the incumbent (and other firms) to invest efficiently in infrastructure and innovation.

References

Acton, J. and I. Vogelsang, 1992, Telephone demand over the Atlantic: Evidence from country-pair data, Journal of Industrial Economics, 40, 305–323.

Armstrong, M., 1997, Mobile telephony in the UK, Regulation Initiative Discussion Paper No.15, London Business School.

Armstrong, M., 1998, Network interconnection in telecommunications, Economic Journal, 108, 545–564.

Armstrong, M., 1999, Competition in the pay-TV market, Journal of the Japanese and International Economies, 13, 257–280.

Armstrong, M., 2001, Access pricing, bypass and universal service, American Economic Review, 91, 297–301.

Armstrong, M., S. Cowan and J. Vickers, 1994, Regulatory reform: Economic analysis and British experience, Cambridge, MA: MIT Press.

Armstrong, M., C. Doyle and J. Vickers, 1996, The access pricing problem: A synthesis, Journal of Industrial Economics, 44, 131–150.

Armstrong, M. and J. Vickers, 1993, Price discrimination, competition and regulation, Journal of Industrial Economics, 41, 335–360.

Armstrong, M. and J. Vickers, 1998, The access pricing problem with deregulation: A note, Journal of Industrial Economics, 46, 115–121.

Armstrong, M. and J. Vickers, 2001, Competitive price discrimination, RAND Journal of Economics, 32, 579–605.

Baumol, W., 1983, Some subtle issues in railroad regulation, International Journal of Transport Economics, 10, 341–355.

Baumol, W., 1999, Having your cake: How to preserve universal-service cross subsidies while facilitating competitive entry, Yale Journal on Regulation, 16, 1–18.

Baumol, W., J. Ordover and R. Willig, 1997, Parity pricing and its critics: A necessary condition for efficiency in the provision of bottleneck services to competitors, Yale Journal on Regulation, 14, 145–164.

Baumol, W. and G. Sidak, 1994a, The pricing of inputs sold to competitors, Yale Journal on Regulation, 11, 171–202.

Baumol, W. and G. Sidak, 1994b, Toward Competition in Local Telephony, Cambridge, MA: MIT Press.

Biglaiser, G. and M. Riordan, 2000, Dynamics of price regulation, RAND Journal of Economics, 31, 744–767.
Carter, M and J. Wright, 1994, Symbiotic production: The case of telecommunication pricing, Review of Industrial Organisation, 9, 365–378.
Carter, M and J. Wright, 1999, Interconnection in network industries, Review of Industrial Organisation, 14, 1–25.
Carter, M and J. Wright, 2001, Asymmetric network interconnection, mimeo.
Cave, M. and M. Donnelly, 1996, The pricing of international telecommunications services by monopoly operators, Information Economics and Policy, 8, 107–123.
Crew, M. and P. Kleindorfer, 1998, Efficient entry, monopoly and the universal service obligation in postal service, Journal of Regulatory Economics, 14, 103–126.
De Bijl, P. and M. Peitz, 2002, Regulation and entry into telecommunications markets, Cambridge, Cambridge University Press.
De Fraja, G., 1999, Regulation and access pricing with asymmetric information, European Economic Review, 43, 109–134.
Dessein, W., 2000, Network competition in nonlinear pricing, mimeo, University of Chicago.
Diamond, P. and J. Mirrlees, 1971, Optimal taxation and public production: I – Production efficiency, American Economic Review, 61, 8–27.
Doane, M., D. Sibley and M. Williams, 1999, Having your cake: How to preserve universal service cross subsidies while facilitating entry (response), Yale Journal on Regulation, 16, 311–326.
Domon, K and K. Kazuharu, 1999, A voluntary subsidy scheme for the accounting rate system in international telecommunications industries, Journal of Regulatory Economics, 16, 151–165.
Economides, N., 1998, The incentive for non-price discrimination by an input monopolist, International Journal of Industrial Organisation, 16, 271–284.
Economides, N. and L. White, 1995, Access and interconnection pricing: How efficient is the efficient component pricing rule? The Antitrust Bulletin, 40, 557–579.
Gans, J., 2001, Regulating private infrastructure investment: Optimal pricing for access to essential facilities, Journal of Regulatory Economics, 20, 167–189.
Gans, J. and S. King, 2000, Mobile network competition, customer ignorance and fixed-to-mobile call prices, Information Economics and Policy, 12, 301–328.
Gans, J. and S. King, 2001, Using 'bill and keep' interconnect agreements to soften network competition, Economics Letters, 71, 413–420.
Gans, J. and P. Williams, 1999, Access regulation and the timing of infrastructure investment, Economic Record, 79, 127–138.
Hahn, J.-H., 2000, Network competition and interconnection with heterogeneous subscribers, mimeo, Keele University.
Hakim, S. and D. Lu, 1993, Monopolistic settlement agreements in international telecommunications, Information Economics and Policy, 5, 147–157.
Hausman, J., 2002, Mobile telephony, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science B. V., 563–604.
Hausman, J. and G. Sidak, 1999, A consumer-welfare approach to the mandatory unbundling of telecommunications networks, Yale Law Journal, 109, 420–505.
Hermalin, B. and M. Katz, 2001, Network interconnection with two-sided user benefits, mimeo, University of California, Berkeley.
Jeon, D.-S., J.-J. Laffont and J. Tirole, 2001, On the receiver pays principle, mimeo, Barcelona: Universidad Pompeu Fabra.
Jorde, T., G. Sidak and D. Teece, 2000, Innovation, investment and unbundling, Yale Journal on Regulation, 17, 1–37.
Kim, J.-Y. and Y. Lim, 2001, An economic analysis of the receiver pays principle, Information Economics and Policy, 13, 231–260.

Klemperer, P., 1995, Competition when consumers have switching costs, Review of Economic Studies, 62, 515–539.
Laffont, J.-J., S. Marcus, P. Rey and J. Tirole, 2001, Internet interconnection and the off-net-cost pricing principle, mimeo, Toulouse.
Laffont, J.-J., P. Rey and J. Tirole, 1998a, Network competition: I. Overview and non-discriminatory pricing, RAND Journal of Economics, 29, 1–37.
Laffont, J.-J., P. Rey and J. Tirole, 1998b, Network competition: II. Price discrimination, RAND Journal of Economics, 29, 38–56.
Laffont, J.-J. and J. Tirole, 1994, Access pricing and competition, European Economic Review, 38, 1673–1710.
Laffont, J.-J. and J. Tirole, 1996, Creating competition through interconnection: Theory and practice, Journal of Regulatory Economics, 10, 227–256.
Laffont, J.-J. and J. Tirole, 2000, Competition in telecommunications, Cambridge, MA: MIT Press.
Lapuerta, C. and W. Tye, 1999, Promoting effective competition through inter-connection policy, Telecommunications Policy, 23, 129–145.
Lee, S.-H. and J. Hamilton, 1999, Using market structure to regulate a vertically integrated monopolist, Journal of Regulatory Economics, 15, 223–248.
Lewis, T. and D. Sappington, 1999, Access pricing with unregulated downstream competition, Information Economics and Policy, 11, 73–100.
Lipsey, R and K. Lancaster, 1956, The general theory of the second best, Review of Economic Studies, 24, 11–32.
Little, I. and J. Wright, 2000, Peering and settlement in the Internet: An economic analysis, Journal of Regulatory Economics, 18, 151–173.
Mandy, D., 2000, Killing the goose that laid the golden egg: Only the data know whether sabotage pays, Journal of Regulatory Economics, 17, 157–172.
Mayer, W., 1981, Theoretical considerations of negotiated tariff agreements, Oxford Economic Papers, 33, 135–153.
Poletti, S. and J. Wright, 2000, Network interconnection with participation constraints, Mimeo, University of Auckland.
Rey, P. and J. Tirole, 1996, A primer on foreclosure, mimeo, IDEI, Toulouse.
Riordan, M., 2002, Universal residential telephone service, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science B. V., 423–473.
Sibley, D. and D. Weisman, 1998, Raising rivals' costs: The entry of an upstream monopolist into downstream markets, Information Economics and Policy, 10, 451–470.
Sidak, G. and D. Spulber, 1996, Deregulatory takings and breach of the regulatory contract, New York University Law Review, 4, 851–999.
Sidak, G. and D. Spulber, 1997a, Deregulatory Takings and the Regulatory Contract, Cambridge: Cambridge University Press.
Sidak, G. and D. Spulber, 1997b, Givings, takings and fallacy of forward-looking costs, New York University Law Review, 5, 1068–1164.
Tirole, J., 1988, The Theory of Industrial Organisation, Cambridge, MA: MIT Press.
Tye, W. and C. Lapuerta, 1996, The economics of pricing network interconnection: Theory and application to the market for telecommunications in New Zealand, Yale Journal on Regulation, 13, 419–500.
Vickers, J., 1995, Competition and regulation in vertically related markets, Review of Economic Studies, 62, 1–17.
Vickers, J., 1996, Market power and inefficiency: A contracts perspective, Oxford Review of Economic Policy, 12, 11–26.
Vickers, J., 1997, Regulation, competition and the structure of prices, Oxford Review of Economic Policy, 13, 15–26.

Weisman, D., 1995, Regulation and the vertically integrated firm: The case of RBOC entry into interLATA long distance, Journal of Regulatory Economics, 8, 249–266.
Willig, R., 1979, The theory of network access pricing, in H. Trebing, ed., Issues in Public Utility Regulation, East Lansing, MI: Michigan State University Press,
Wright, J., 1999, International Telecommunications, settlement rates and the FCC, Journal of Regulatory Economics, 15, 267–291.
Wright, J., 2002, Access pricing under competition: An application to cellular networks, Journal of Industrial Economics, forthcoming.
Yun, K., H. Choi and B. Ahan, 1997, The accounting revenue division in international telecommunications: Conflicts and inefficiencies, Information Economics and Policy, 9, 71–92.

Chapter 9

INTERCONNECTION PRACTICES

ELI M. NOAM*

Columbia University and Columbia Institute for Tele-Information

Contents

* The author gratefully acknowledges help by Robert Atkinson and Kenneth Carter.

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Interconnection as the key policy tool of telecommunications

For more than a century, telecommunications around the world followed a classic model: a national monopoly owned or controlled by the state, centrally managed and providing a common public network. By their very nature and tradition, these networks provide a small number of standardised and nation-wide services, carefully planned, methodically executed, and universally distributed. But over the past three decades, first in the United States and subsequently in much of the developed world, the forces of centrifugalism began to unravel this traditional system. The driving force behind the restructuring of telecommunications was the shift toward an information-based economy, which resulted in the rapid growth of telecommunications as the medium for the electronic transmission of information.

The tension between the convergent forces of technology and the centrifugal forces of business competition is most pronounced on the front where they intersect: the rules of interconnection of the multiple hardware and software networks and their integration into a whole. As the various discrete networks grew, they had to inter-operate. In the networks of networks, their *interconnection* became critical. Control of interconnection by any entity, whether by government or by a private firm, became the key to the control of the telecommunications system and its market structure.

The regulation of wholesale interconnection has therefore emerged as the paramount tool of regulation and is likely to remain so into the reasonably foreseeable future, replacing the regulation of telecommunications retail pricing of network operators, or of the entry of competitors.

The term 'interconnection' is defined by the International Telecommunication Union as:

> "The commercial and technical arrangements under which service providers connect their equipment, networks and services to enable customers to have access to the customers, services and networks of other service providers."

The most traditional form of interconnection has been *parallel* or *co-operative* interconnection. In that arrangement, dominant carriers link up with carriers similarly dominant in other regions. Their relation is that of partners and 2-way correspondents; they jointly extend network externalities to their customers and often raise their prices in a joint maximisation strategy. This cartel type of interconnection is in decline due to its inability to maintain control over entry.

The second classic interconnection arrangement is *vertical*, between a provider that possesses market power in one stage of the transmission chain and another provider that requires use of the bottleneck in order to provide service. An example would be a long-distance company interconnecting into a local exchange carrier. This type of interconnection has been contentious since the early

days of telecommunications. It has been studied and analysed over the years, but new permutations keep emerging. For example, the interconnection by Internet service providers into cable TV networks, or of fixed networks into mobile carriers.

More recently a third type of arrangement has been taking centre stage, that of a *horizontal* interconnection, in which competitors for the same markets and customers link up with each other. In the past this situation was suppressed by the stronger of the two parties, often with the support of government, sometimes in return for the fulfilment of a number of social obligations of redistribution. In other cases, the horizontal participants were kept apart from each other by technology and regulation, as for cable television and telecom networks. Today, many governments enable and even promote the emergence of such horizontal interconnection.

The term 'interconnection' covers a wide matrix of relations. On the physical level of transmission *conduits* they include linkages within and among various types of entities and industries:

1. Incumbent and new local telephone companies
2. Traditional and new long-distance carriers
3. Mobile and radio carriers, including their access to spectrum
4. Domestic and international carriers
5. Dedicated 'private networks' of organisations and user groups
6. Computer local area and wide area networks, in particular the components of the Internet, such as backbones
7. Telephone, computer, and video equipment
8. Cable television, broadcast and satellite networks

On the higher levels of *applications* and *content*, interconnection becomes an issue for entities such as:

1. Internet service providers
2. Enhanced (value-added) service providers
3. Data and information providers
4. Video program channels.

On a *geographic* level, interconnection issues cross national boundaries and involve the carriers, service providers, and national policy makers of many countries. The *direction* of flows is another dimension: terminating vs. originating traffic, and one-way vs. two-way directionality.

Given the multitude of entities, their points of intersection are numerous and growing, so is the number of disputes and issues – technical, financial, operational, regulatory, international, and content-wise. Their common thread is the transfer of information streams from network facilities of one communications entity to those of another.

1.1. Why regulate interconnection

1.1.1. Anti-monopoly rationale

There are two major explanations for government's role, coexisting uneasily. The primary explanation for a governmental role in assuring interconnection is market power. It starts with two assumptions: (1) that telecommunications are a service essential to society and economy and, (2) that monopoly provision is undesirable. Given the incumbent's head start of a full century, economies of scale and scope, and the positive externalities of its reach, a new entrant cannot hope, it is argued, to succeed as a stand-alone entity. Yet the entrant must reach the customers of the incumbent and, in turn, be reachable by them. Thus, if one wants to encourage competition to a strong incumbent, one must accompany it with an assurance of interconnection. And if the survival of fledgling competition is at stake, this rationale is readily expanded to justify interconnection on terms that are favourable for an entrant as an 'infant' period.

The flip side of the anti-monopoly rationale is that a carrier without market power would owe no interconnection to anybody. This means an asymmetrical arrangement among carriers. This, in turn creates instability. If interconnection rights vanish with bottleneck power, the determination of that point is fiercely fought over. The question, after all, is not an easy one to answer conceptually or empirically, and it may vary by location, service and customer class.

1.1.2. Transaction cost rationale

The other major rationale for the regulation of interconnection might be called the 'transaction cost' explanation. This view centres on the positive externalities of networks. Interconnection is designed to provide an element of integration to the increasingly disparate network environment. Information flows across numerous pathways, in a chain of transmission involving half a dozen carriers. Indeed, with packet-switched communication, which is the mainstay of much of Internet communications, information between two points may travel simultaneously over a wide variety of paths. In such an environment, interconnection rules are a transaction-cost reducing arrangement, and as such are similar to legally imposed arrangements aimed at reducing transaction cost in other parts of the economy. The interconnection rules may limit some freedom of negotiation, but they also facilitate commerce and transactions. They establish symmetry in the treatment of various carriers, and eliminate continuous market power tests.

1.2. Regulation of interconnection and unbundling in a competitive market

The historic experience with interconnection around the world shows that interconnection is not made available freely by an incumbent to its competitors. Nor is the claim to interconnection as a right given up voluntarily by new entrants once

competition emerges. On the other hand, interconnection is voluntarily initiated by collaborating and non-competing carriers, such as those of different countries.

Often, the terms of interconnection are left nominally or initially to the parties' negotiation. Yet regulatory intervention is frequent where there is an asymmetry in bargaining strength and in the urgency for interconnection, which is usually the case. Even where formal regulatory intervention does not take place, the negotiations are shaped by the expectations of what the regulator's decisions would be. Those decisions, in turn, depend on fundamental policy priorities.

As a matter of empirical fact, interconnection is regulated everywhere where-competitive telecommunications exist. Even in New Zealand, which was supposedly without any telecommunications regulation, the courts of law and their interpretation of the statutes of general competition regulated interconnection. The difference is institutional – a general regulatory body vs. a specialised agency – and it is not clear whether their substantive policy decisions would be fundamentally different or better.

Today the antimonopoly and the transaction cost views coexist uneasily, but they differ in their perspectives of the future. In the antimonopoly view, the regulation of interconnection is an essentially transitional task that will fade away with the emergence of real competition. Interconnection regulation would decline over time. In contrast, the transaction cost rationale comes to the opposite conclusion. As open entry permits more and more carriers to offer services, the need for basic rules for their interaction becomes increasingly important if the overall network infrastructure is not to fragment into incompatible network parts. The antimonopoly view is asymmetric, requiring interconnection by large carriers but not by their competitors. In contrast, the transaction cost view is symmetrical, applying interconnection to all carriers.

2. Interconnection as a tool for the creation of monopoly: the U.S. experience

Interconnection is not a new issue but goes back over a full century. Control over interconnection was used to establish the monopoly system. It was later used in the second stage of interconnection policy to introduce competition. In its third and present phase, interconnection policy is increasingly used to promote competition[1].

In the United States, the initial monopoly was based on patents rather than regulation. Once the basic Bell patents expired in the 1890s, independent competitors entered, especially in rural districts and in central business districts. The Bell Company's initial policy was to refuse interconnection to the independents. Its strategy centred on the control of interconnection: of equipment into their network, of rival local networks into the Bell local networks, and of rival

[1] Noam (2001).

networks to the Bell long-distance system. Interconnection was granted, if at all, through a contractual agreement, because this allowed AT&T to exercise its substantial bargaining power. In contrast, the new entrants preferred interconnection as a matter of legal *right* and sought mandatory interconnection supervised by the state. This scenario is classic, and it has repeated itself in recent times around the world.

In Europe, too, early control over interconnection was used to establish monopoly. Especially in Sweden, Norway and Britain, competitors were initially successful, but monopoly soon took over through control over interconnection. No regulatory counter-force protected competitors against the dominant operator, which was government itself.

In the United States, interconnection became regulated. In 1904 a federal court upheld the powers of a state to mandate the interconnection of rival networks. By 1915 more than thirty states had such requirements[2], and all still do so today.

Soon, however, interconnection was used to stabilise a cartel. Several independent telephone companies brought federal antitrust complaints against AT&T, based on its refusal to offer interconnection, and they were joined by the Justice Department in 1913. Under pressure, AT&T accepted an agreement known as the *Kingsbury Commitment* which granted the independents interconnection in return for joining an AT&T led cartel. A system of de facto exclusive franchises emerged in which only one telephone company served any particular area. That company was protected from rival entry because no rival had a right to interconnection. Thus the independents now had an interest in an effective Bell system as an interconnecting agent, technology driver, standard setter, and cartel enforcer. As a result, the telephone industry moved from extensive competition to extensive oligopolistic co-operation.

This historic episode illustrates that the creation of interconnection, by itself, does not necessarily bring about competition, and can in fact lead to cartel co-operation that turns new entrants into complements rather than competitors[3]. Thus, interconnection does not assure competition, but the lack of such interconnection has historically prevented its emergence. Interconnection has been a necessary but not sufficient condition for competitive telecommunications.

3. Interconnection as a tool for competitive entry

For over seventy years AT&T's control over interconnection provided it with the tools to establish a monopoly shared with small 'independent carriers.' However, the power that AT&T had ceded to the government to regulate such interconnection had the potential to turn against it. This began to happen in the 1960s,

[2] Gabel and Weiman (1994).
[3] Mueller (1988).

and interconnection now became a tool for destabilising AT&T. Interconnection policy moved into its second stage, that of opening of markets.

The first step was to interconnect customer terminal equipment. The prohibition of such attachment has enabled AT&T to shift earnings to the unregulated manufacturing activities and away from the profit-regulated network services, and to shift costs in the other direction. The two key decisions[4] were Hush-A-Phone (1956) and Carterfone (1966) which allowed competitive equipment to be owned by customers and to be connected to the network.

When it came to the next chapter, the interconnection of long-distance networks, the United States began to move into uncharted water. The availability of microwave transmission equipment after the Second World War drastically lowered economic and technological entry barriers for long distance communication. In 1969, one company, Microwave Communications, Inc. (MCI) won a court ruling[5] against a reluctant FCC and an adamant AT&T to provide private line service[6]. This set the stage for a battle over long-distance interconnection. MCI soon wanted to expand into generally available public switched service. To do so successfully, it needed to interconnect its long haul private lines with AT&T's local networks in order to connect its subscribers to non-subscribers, and vice-versa, and it won a court appeal against an unfavourable FCC ruling in the Execunet decision (1978).

Interconnected competition had cataclysmic effect on the U.S. telecommunications industry. Eventually it led to the break-up of the world's largest telecommunications company, AT&T, based on an antitrust case, which asserted that incumbent had used unfair practices to suppress its competitors, especially through discriminatory interconnection practices. The government's lawsuit resulted, after a 1982 consent decree, in the most massive corporate reorganisation in business history.

3.1. Reforming access charges

The break-up of the Bell System left U.S. regulators scrambling to create a new system of interconnection prices, balancing the efficiency goals of economics with the redistributional aims of social policy and inter-regional politics. The main questions were whether interconnection prices would be usage-sensitive or flat rate, and what their magnitude would be. Each answer had major implications for some industry segments, geographic region or user group. In 1982, the FCC approved an access charge plan[7] adopting the economists' preferences

[4] Hush-A-Phone Corp. v. United States, 238 F.2d 266 (District of Columbia Circuit Court 1956) and Carter v. AT&T, 250 F. Supp. 188 (Northern District Court of Texas 1966).
[5] MCI Telecommunications Corp. v. FCC, 561 F.2d 365 (District of Columbia Court 1977), cert. denied, 434 U.S. 1040.
[6] FCC (1969).
[7] FCC (1982).

for flat charges to recover fixed costs, instead of usage-sensitive ones which would distort usage or encourage uneconomic 'bypass' of customers to other carriers. But the FCC also kept a strong element of redistribution and subsidy in order to maintain widespread telephone connectivity (universal service). The access charge plan included a flat monthly per line charge on users as well as a variable one on long-distance carriers. The share of the costs thus recovered for interstate usage was 25 percent of total line costs, maintaining transfer from long distance to local service. This basic plan was adjusted repeatedly.

The Telecommunications Act (1996)[8] changed many aspects of competition and regulation. A section entitled 'Interconnection Requirements' was included in the law. The Bell companies, eager to enter the long-distance market, had in return to provide interconnection to new entrants, unbundle their network, allow the resale of their services by competitors, and provide for number portability[9].

Following the mandate of the act the FCC took a further step to flat rate access charges, and removed some implicit universal service transfers in favour of explicit support mechanisms. Subscriber line charges – the flat fees paid by customer directly – were raised, and the incumbent local exchange companies (ILECs) were authorised to assess another flat charge, a 'pre-subscribed inter-exchange carrier charge' (PICC) on the long distance carrier chosen by the end user[10]. After further modification towards a flat rate system, by the year 2002, per minute interconnection charges between incumbent LECs (CLECs) and competitive LECs typically ranged from 0.27 cents to 0.55 cents per minute, and were coming down toward 0.10 cents. Per minute termination rates for different types of wireless, paging, and inter-exchange service were roughly 0.30 cents, 0.40 cents, 0.55 cents, respectively. These numbers were declining, and the spread between them was narrowing, thus reducing the incentives for arbitrage. The per-minute charges for enhanced services, including Internet telephony, remained zero, leading to suggestions that all interconnection access charges be abolished (bill-and-keep), which would complete the transition from usage-surcharge to flat-rate charges. Universal service subsides were maintained and partly supported by a surcharge on users' phone bills, which were usage sensitive in several of their components.

4. Interconnection as a tool for protecting competition

4.1. Local competition

Competition in local infrastructure is the toughest challenge for entrants, given the large investments needed. It also the key to a 'level-playing field' for the other telecom services. Since virtually every communications flow ultimately has some

[8] Telecommunications Act of 1996, 47 U.S.C. § 101 et seq., Pub. L. No. 104-104, 110 Stat. 56.
[9] 47 U.S.C. § 271 (1996).
[10] FCC (1997a).

local component, a monopoly in that segment affects all of telecommunications. It is therefore not surprising that policy makers, once they had embarked on a pro-competition strategy, were eager to remove the last bottleneck. Even if deregulation-minded, an active interconnection policy became the tool to accelerate local entry.

In the United States, in the mid-1980s a second wave of competitive entry into telecommunications began. Private line dedicated local service was first approved in the U.S. by the New York Public Service Commission in 1985 for Teleport Communications. New York was also the first state to subsequently permit competitive switched local exchange service. By 1995 most major states had approved competitive local entry. The federal 1996 Telecommunications Act extended this across the rest of the country.

The viability of local competition rises and falls with interconnection. The critical issue surrounding the entrant competitive local exchange companies (CLECs) is whether they can cost-effectively interconnect to the incumbents' – the ILECs'—network under the same conditions provided by the integrated incumbent to its own operations 'comparably efficient interconnection' (CEI). Whether CEI needed to be offered by ILECs to their rivals, and what CEI actually means in technical and economic terms becomes a subject of intense struggle.

One form of CEI requirement is the placement CLEC of cables and equipment inside the ILEC's central office, known as 'physical collocation.' This was resisted by the incumbent as intrusive to their property rights. An alternative is a handoff at an outside meetpoint such as a manhole, from whence the ILEC would carry the CLEC's traffic to its central office on its own facilities and charge for this transport element and associated equipment at retail rates. This alternative is more expensive for the CLECs. A third arrangement is 'virtual collocation,' either with physical handoff taking place outside the LEC, or inside the central office, with the LEC owning all the cable and termination equipment, but with the charges being equivalent to those the CLEC would have incurred with 'physical collocation.' A fourth option is for CEI arrangements to be freely negotiated by the parties.

In the United States, the local entrant Teleport Communications gained in 1987 approval from the New York Public Service Commission for physical collocation of private line services. This was soon expanded to switched public service. Once several other states had liberalised local and physical or virtual collocation entry the FCC had enough 'state cover' to establish national rules for collection for switched services in 1993.

Such interconnection is available not only to CLECs but also to inter-exchange carriers and even to some end-users. There is no reciprocity for physical collocation. The rules on collocation have been attacked as a regulatory taking of private property without compensation. Whether they are or not is a matter of constitutional interpretation but they are certainly not 'deregulatory' in the sense of non-intervention.

The 1996 Telecommunications Act modified FCC and state rules, requiring each telecommunications carrier to interconnect directly or indirectly with other telecommunications carriers and for ILECs to unbundle. Interconnection agreements need to be approved by the respective state utility commission. Together with the FCC's subsequent rules of implementation, a set of strong interconnection rights had been established for access into most local networks and incentives for the Bell Companies were created to open their local networks to interconnection if they wanted approval to enter most long-distance services.

4.2. Unbundling

Interconnection is fairly meaningless without reference to where interconnection would physically take place. If an incumbent network offers an entrant interconnection at a far-off point, little is resolved. In consequence, interconnection points are established at various levels of a network, thereby 'unbundling' it.

Unbundling requirements, too, are significant regulatory interventions. They regulate in order to deregulate. They aim to create viable competition, especially in the early phases of competitive entry when entrants are likely to be weak. There are several advantages to an incumbent in bundling:

1. Bundling forces a competitor to buy unneeded services to get needed ones, thus raising the competitors' operating costs. Bundling is a tying action. Market power in one component can be extended by bundling to a component where market power does not exist. It might shift profits and hide them from regulators.
2. Bundling on the retail level against a competitor, using a monopoly input on the wholesale level, permits a price squeeze.
3. Bundling permits price discrimination based on utilisation among users. IBM, for example, used to bundle its machines with a requirement contract for punch cards, which were priced above cost. This allowed it to charge high-volume users of its machines more than low-volume users[11].

In a competitive environment, markets will determine the extent of unbundling that firms will offer. Bundling is primarily a problem where market power exists. Bundling raises entry barriers. Hence, where interconnection was actively encouraged, unbundling requirements were usually adopted.

To the incumbent, the greatest problem of unbundling, beyond the loss of the tools described above, is that it is prevented from being the exclusive or primary beneficiary of superior efficiencies or talents in a particular market segment, because it cannot shield those efficiencies from access by its own competitors. This is particularly a problem when unbundling and interconnection are asymmetric,

[11] IBM vs. U.S. (1936).

leaving a competitor with access to an incumbent's superior elements, while not having to grant those in reverse.

Suppose one has been able to unbundle the monopolistic parts of the network. Do the monopolistic network elements need unbundling from each other? Where the monopoly is otherwise unregulated, this seems pointless, because bundling unneeded service elements is just a way to raise prices, which could be accomplished in other ways, too. But if regulation limits profits or prices on each element, to require a competitor to take even unneeded ones is a way to raise the price. In such cases, regulated unbundling is likely.

In 1996, following the requirements of the Telecommunications Act, the FCC issued a voluminous order[12]. It prescribed minimum points of interconnection as well as adopted a list of unbundled network elements (UNEs) that the incumbent LEC must make available for competitors. These were seven: network interface devices, local loops, local and tandem switches, interoffice transmission facilities, signalling and call-related database facilities, operations support systems and information, and operator and directory assistance facilities. Incumbent LECs were required to provide equal and non-discriminatory access to these elements in a manner that allowed the entrant to combine such elements as they chose and prohibited incumbent LECs from imposing any restrictions on the use of these elements. In 1999, the FCC also unbundled the high frequency range on the copper loop, thereby making it available to rival digital subscriber loop (DSL) providers as a separate network element. It also dropped operators and directory services, unbundled sub-loops, and established periodic reviews to carve out geographic regions and network elements that had become open and where entrants required no regulated UNEs.

Unbundled network elements are priced in the U.S. using forward-looking long-run incremental cost pricing principles (TELRIC) discussed below. The actual calculations are based on engineering models developed initially for the calculation of universal service cost allocations. These models were then applied by the states and diverged widely, ranging from $3 for a local loop in one state to almost $30 in another.

The unbundling of the local loop facilitated entry into sub-markets such as local and tandem switching, and into the transport segments into between. It is harder to determine the impact on facilities-based competitive entry into the local loop itself. On the one hand, it helped a competitor to stage a gradual move in the direction of the user; on the other hand, it also made it possible to enter local competition without the heavy investments in local loop infrastructure. On the whole, it provided more flexibility in providing local service.

By the year 2000, the concept of network unbundling was firmly established in the United States, the European Union, Japan, and in the regulatory principles of the WTO. Though unbundling is a significant regulatory intervention, it had

[12] FCC (1996).

become a key tool for governments pursuing pro-competitive policies. Yet it was also possible to anticipate its reduction. With competition, one could expect the unbundled elements to shrivel down to the last and most expensive to enter–the last part of call termination on the local loop beyond the switch, or even further downstream to the sub-loop. Unbundling facilitated such transition, though at the cost of greater operational and procedural complexity.

4.3. Quality

Society depends more and more on the availability of electronic communications. The World Wide Web is an example. User requirements keep increasing rapidly. In consequence, demands on service quality increase because failure becomes more costly. In a transmission sequence of multiple carriers, a signal quality will not normally be better than the 'weakest link.' Hence, a bottleneck carrier with inferior quality could obviate the efforts of other carriers for higher quality. They might lower their quality to the lowest common denominator. Thus, overall quality would decline.

This and similar reasons led to the fear that a decentralised competitive environment would lead to service degradation. On the empirical level, measuring the quality service is quite complex and even more so across carriers. The term 'quality' has many dimensions, and measuring problems abound. On the conceptual level, economic analysis does not provide unambiguous answers on what to expect to happen to quality as competition emerges. A monopoly need not compete for users by offering superior quality, but competition might lead to a low price, low quality equilibrium. Regulatory incentives might lead a monopoly to overcapitalisation and above-equilibrium quality. A more competitive regime may well reduce such overcapitalisation and lead to an economically more efficient, but lower-quality system.

4.4. Cable television interconnection

Some cable television operators, starting in Britain in the 1980s, began to offer local telephone service. These calls are carried on the cable television companies' separate or integrated lines and are then usually interconnected with the networks of local exchange and long-distance carriers.

In the emerging network of networks the distinction between telephone and cable lines blurs, since upgraded telephone lines are able to carry broadband services including video and high-speed Internet services, while cable lines connected to switching and routing equipment are able to perform traditional telephone functions, as well as provide Internet transmission[13].

[13] Noam (1994a).

This technical and business convergence raises the question of regulatory convergence. It is relatively easy to deal with cable interconnection into phone networks. Here, cable companies are treated in the same way as any other local CLEC competitors. But it is more difficult to deal with other types of services. A special problem is the issue of Internet connectivity, specifically whether a cable television firm providing Internet-capable transmission may provide preferential terms to some Internet service providers (ISPs) and web portals. Phone companies are restricted from similar discrimination by common carrier obligations. This issue burst to the fore in 1998 when AT&T acquired the largest and third largest American cable TV companies, TCI and Media One. AT&T's strategic aim was to benefit from its economies of scope by offering alternative local telephone service bundled with long-distance, cable TV, wireless, Internet services, and its own cable-based ISP @Home and the web portal Excite. @Home and Excite received preferential access to AT&T cable customers. AOL, the largest of ISPs initiated a fierce regulatory fight against AT&T through all levels of government.

The rival ISPs demanded an equal access arrangement to cable TV networks. After much legal wrangling, an appellate court classified in 2000 the Internet access provided by a CATV company as a 'telecommunications service' subject to federal telecom regulation. Conversely, the ICC in 2002 was poised to define such access a an "information service" which meant that cable companies in their Internet provisioning were subject to none of the regulations aimed at telephone companies, such as unbundling and interconnection requirements. This set the stage for a process of regulatory convergence of mass media and telecom carriers, attendant with a full set of issues of great conceptual, technical, financial and political complexity.

Paralleling these developments was the approval process for the merger of the media giants Time Warner and AOL. In 2001 the anti-trust agency FTC and the communications agency FCC conditioned their agreement to the merger to the companies committing themselves to meet several condition. Time Warner would have to provide high-speed cable access to at least three competitors. Where agreements were not reached, the FTC could appoint a trustee to negotiate an access agreement on AOL Time Warner's behalf. Access agreements were subject to a 'most favoured nation' clause if other ISPs get subsequently a better deal. The company could not discriminate against content provider by other ISPs. And AOL had to open its vast instant messaging community to access by other IM providers for the next version of advanced IM. These were fairly far-reaching conditions.

4.5. Mobile interconnection

Cellular telephony was a major improvement in mobile communication because it can reuse the same frequencies in multiple geographical areas, called cells. Each cell is connected to central serving points and from there the traffic may be routed

into a LEC, an IXC, or another carrier. The wireless carrier has to pay for its part of the traffic that is routed onto the landline fixed carrier's transmission facilities.

In the U.S., the FCC left the terms and conditions of mobile-fixed interconnection to the states, and to direct negotiations between the parties. Interconnection, in some instances, state regulation fixed many of these terms with a 'standard contract.'

The flow of traffic from mobile systems to fixed-link systems is significantly higher than the reverse flow. Thus, the balance-of-traffic is skewed. This has led to asymmetric interconnection charges that are higher for call imports by mobile networks than for exports[14]. The issue pivots on who pays for an incoming wireless call. In most countries the calling party pays (CPP). In the U.S., it has long been the receiving party that pays (RPP). CPP makes it cheaper to initiate calls to mobile phones and to use prepaid cards, thus increasing volume. Yet, in a CPP system, LEC customers might be confronted with bills for calls they imagined to be local but which were, in fact, calls to a mobile customer with higher per-minute charges. The CPP system leaves the caller to the pricing policy of the mobile operator, with no competitors available to reach the mobile user, because that choice is made by the non-paying party, the recipient, who might receive low prices to let those who call him be charged at a high rate. According to an ITU study, for European fixed-to-mobile network interconnection, the average per minute interconnection charge fixed-mobile was $0.21 per minute in 2000. The charge in the opposite direction, from mobile into fixed, was much lower, by a ratio of 20:1[15]. In contrast, in the U.S. with its RPP system, per minute access charges from fixed to mobile were only 0.3 cents per minute. These dynamics and numbers have led to price regulations of mobile termination. However, as mobile connectivity becomes pervasive and competitive, mobile carriers may offer flat rated mobile termination and origination arrangements that could reduce this problem.

4.6. Internet interconnection

Interconnection issues have also emerged for the Internet. Its dynamics provide insights for the question whether interconnection would be offered in a competitive market, and on what terms. Since the Internet is a loose federation of autonomous networks, interconnection and access issues abound. The first phase of the Internet in the U.S. was government dominated through provision of the first backbones, initially by the Defence Department ARPANET, then by the NSFNET backbone. In time, this created problems in accommodating other backbone networks and routing packets when different backbone alternatives became available. There were technical protocol problems as well as economic issues involving financial settlements among networks and ISPs. Interconnection

[14] Cave (1994).
[15] Melody and Samarajiva (2000).

nodes are known as Network Access Points or NAPs. For a single fee an ISP can access, at the NAP, the other backbones present at that location. Several backbones formed their own commercial interconnection point, the Commercial Internet Exchange (CIX) in Santa Clara in 1991, and agreed on a settlement-free traffic exchange. CIX could not dominate interconnection because rival interconnection arrangements existed. On the East Coast, the Metropolitan Area Exchange-East (MAE-East), owned by the UUNET (and acquired by WorldCom), was created to provide bilateral interconnection arrangements among major backbones. No uniform multilateral agreement or settlement payments exist. Instead, the backbones negotiate agreements with each other.

Even with such competition in interconnection, the existing system of multichannel-backbone 'public' meeting point often exhibited bottleneck characteristics. Many ISPs and backbones therefore shifted to direct 'peering' and regional arrangements.

Similar arrangements exist in many countries. In some countries the function of NAPs is not only technical but also regulatory. NAPs are used for the control of content in countries such as China and Saudi Arabia which limit their citizens' access to foreign information. Thus interconnection can become a competitive service, but it can also be a tool for control.

Key features of peering arrangements are that each Tier 1 'core' ISP has a separate interconnection arrangement with each of the other Tier 1 ISPs, and that such ISP accepts traffic destined to its customers but not to another Tier 1 ISP's customer. This reduces competition. ISPs in peering relationships most typically engaged in 'bill-and-keep' rather than make settlement payments to one another[16]. However, it is difficult for any ISP to assure that its partners do not 'dump' traffic to it as a default route[17]. If settlements (i.e., payments) do not take place, each ISP has an incentive to use the other's backbone rather than its own, known as the 'hot potato' strategy. Peering arrangements of large backbones with smaller ISPs require financial payments, and have at times been used to put financial pressure on smaller competitors and to reduce arbitrage.

Internet interconnectivity is a case of largely unregulated service providers that are partly competing and partly collaborating. Do they voluntarily provide interconnection to each other? The answer is yes. They do so because of the externality advantages to their members in having a larger number of network participants. Several models of interconnection have emerged, including provision by third parties who offer interconnection as their business. Thus competition in the operation of interconnection is likely. On the other hand, the interconnection arrangement among backbone providers also demonstrates their common incentives to eliminate arbitrage by resellers and to restrict smaller competitors.

[16] Lehr and Weiss (1996).
[17] Baake and Wichmann (1999).

Thus interconnection, as an essential service, is subject to similar dynamics as other important activities. Its distribution becomes subject to attempts at raising entry barriers by major industry players seeking to stabilise an industry and pricing structure whose fixed costs are high and marginal costs low. Here, too, control over interconnection becomes control over the market.

5. Pricing and pricing wars

Interconnection comes with a variety of controversies, none more contentious than its price. The setting of interconnection charges can be used as a tool by regulators to finance unrelated policy goals, by incumbents to frustrate competition, and by entrants to gain a subsidy. The challenge for regulators is to set prices for an intermediate good – interconnection – in a way that encourages an efficient competitive entry and avoids an inefficient one. Setting interconnection rates provides wide margins for economics and politics. These prices can determine industry structure and network architecture[18] range of pricing models exists and several will be described in the following.

5.1. Regulated pricing of interconnection

5.1.1. Zero-charge (bill-and-keep) and lump-sum payments

Two networks might agree to a zero-charge where traffic and costs between the entrant and the incumbent are balanced, and where it therefore would be administratively easier to impose no charge. Since costs are not passed on, each carrier has an economic incentive to increase the efficiency of its own network. At the same time, each firm will also try to maximise its outgoing calls in relation to the incoming traffic to divert its traffic to the other carrier as soon as possible, and to discourage usage of its own network. Bill-and-keep also has problems when originating and terminating usage proves unbalanced. In that case, carriers with disproportionately large originating usage will get to keep the most of the revenues, even as they impose the largest cost border on carrier interconnecting into them. A bill-and-keeps arrangement was the initial arrangement among commercial Internet service providers each backbone provider trying to unload its traffic and associated costs as quickly as possible on its partners, and to provide them with poor service to discourage them from doing the same. Bill-and-keep was advocated as an efficient regime for interconnection pricing[19]. For interconnected calls, the calling party would bear the cost of delivery to the called party's central office, and would not have to pay termination from there[20].

[18] Vogelsang and Mitchell (1997).
[19] Brock (1995).
[20] deGraba (2001).

This system would reduce the problem of local terminating monopoly by eliminating the incentive to charge high interconnection prices, since none exist. It also eliminates the discrepancy in treatment of access charges, in the U.S., between ISPs and IXCs.

A related system is to charge a fixed (lump) payment for access by a carrier. Such a fee can vary with the capacity provided. Lump sum fees do not affect the marginal behaviour. But they, too, lead to wasteful usage where marginal costs are non-zero, and they are advantageous to the heaviest of users. At the same time, a high lump-sum charge could lead to incentives for interconnectors to seek a bypass solely to avoid the payment.

5.1.2. Average cost pricing

All cost-based pricing methods have the administrative advantage that they require only information about the providing network and not about users, uses, and interconnectors. Interconnection prices can be set at the average cost of providing such interconnection. The average cost can be based on actual historic or on a hypothetical future-looking costs. This approach incurs therefore many of the same practical problems associated with marginal cost pricing, (discussed below) though it is generally easier to determine total and average costs than marginal ones. The problem with the average costs approach is that they only represent the mean cost of capacity usage. They, therefore, do not reflect cost variations across a given time period, notably cost difference between peak and non-peak usage. Average cost in a capital-intensive industry will also be usually above marginal cost, and such a price would deter entry by interconnection. Another problem is the conceptual and practical application of average cost pricing to multi-product outputs.

5.1.3. Fully distributed cost pricing and two-part tariffs

Fully distributed cost pricing (FDC) tries to combine the economic incentives of marginal cost pricing with a way to cover the fixed costs. In theory, it combines all costs common—fixed and incremental—and allocates them to different services—such as to local and long-distance; residential and business—according to a formula. The cost per unit for service related items is usually based on telephone message minutes or message minute miles. For many years, FDC was the dominant method of regulation.

As with average cost, a basic defect of fully distributed cost is that it exceeds marginal costs, and therefore fails to measure the costs causally imposed[21]. This may encourage entry by firms whose long run marginal costs are higher than those of the incumbent[22]. Another problem is that the allocation of the joint costs

[21] Kahn (1988).
[22] Braeutigam (1980).

of an incumbent in a multiproduct setting is subjective. This leads to protracted disputes over the appropriate allocation of joint and common costs.

A *two-part tariff* combines a flat and a variable charge. It can be composed of a variable usage charge on top of a flat capacity charge. The former is supposed to recover all variable costs and is charged on a per-unit basis. The latter distributes the fixed costs to customers according to their respective basic requirements. Though not using this terminology, this approach has been long used in the U.S. for local access, with a combination of flat and variable charges.

5.1.4. Price caps

Interconnection prices can also be set by *price caps*. In principle, such cap can be tacked on to any pricing scheme except where such a price is zero. A given interconnection price, however derived, is then indexed to inflation, productivity expectations, and other factors. A high productivity factor would in time lead to low prices. A price cap system provides incentives for cost reduction, because a carrier would gain by cutting cost. The productivity factor might turn out to be too generous, or too harsh for an incumbent LEC. It would be subject to periodic recalibration (with its associated regulatory fight over costs and profits) which means that profit measurements (revenue minus cost) do not really disappear. But in the short term a price cap system is a quick way to proceed, due to its gradualism and simplicity.

Nothing in telecommunications regulation stays simple for long, however. Soon, disputes arise over what is included in the 'basket' that is subject to price caps, whether there should be different caps to reflect different production trends of various network elements and services, whether consumers should set a 'retail cap' different from other companies, and whether new networks services and technologies should be excluded[23].

5.1.5. Ramsey pricing

This pricing rule recommends that when marginal cost pricing will not recover total cost, a price discrimination is instituted in which each customer class is charged a price inversely proportional to its demand elasticity. The basic intuition behind Ramsey prices is that to allocate fixed costs among all customers, the heaviest burden is put on those customers who want service badly enough so that their behaviour will be least affected[24]. This welfare optimising principle is also known as the *inverse-elasticity rule*.

[23] Laffont and Tirole (1996).
[24] Baumol and Bradford (1970).

The problems associated with the Ramsey pricing are several. On the policy level, it means that customers with fewer options, often smaller users, would be charged highest, hardly a popular undertaking given that it is the reverse of traditional charging system under which such residential users tend to get a price discount. Furthermore, Ramsey pricing requires the prevention of arbitrage, i.e., significant regulatory monitoring and enforcement, because such arbitrage would destroy the ability to price differentiate. Also, the informational requirement on regulators to set these different prices correctly can be prohibitive. Not only must marginal costs be known, but information on the elasticities for different customer types must be available as well. In most cases, information is incomplete or asymmetric[25]. This situation can be improved somewhat through the creation of proper incentives on the network provider through 'global price caps'[26].

5.1.6. Wholesale pricing

One can approach the question of pricing interconnection differently. Interconnection charges cannot be separated from the more general issue of appropriate telephone rates. One can think of an interconnecting network as a large user. Should such a user pay a rate that is different for the services offered to other large users by the incumbent carrier?

Treating competitive networks as each other's large customers greatly simplifies administrative arrangements. However, it also creates incentive for incumbents to squeeze their competitors through high conveyance prices, to charge usage prices for services whose costs are not traffic-sensitive, and to bundle necessary elements with unnecessary ones. To compete, an entrant needs a price below the retail level for the final service. The key to resale is its price. Or, more precisely, the extent of its discount over retail price. The absolute level of interconnection charge, by itself, is not germane to whether or not a non-integrated entrant is able to compete with an incumbent, as long as the incumbent has to charge itself the same price[27]. Rather, the ability to compete is based on the margin between the interconnection charge and the incumbent's final prices, i.e. in the avoidance of a *vertical squeeze*. A major problem with the wholesale price approach is that for reasons of public policy, many retail prices in telecommunications are kept low, such as for residential and rural customers. Hence, even a cost-based wholesale price might not be low enough to permit the interconnector a retail profit.

On the other hand, a large discount for wholesale prices over retail might require the retail customers of the incumbent to subsidise the entrant and its customers. This would, in turn:

[25] Laffont and Tirole (1994).
[26] For more detail, see Mark Armstrong's contribution in this *Handbook*.
[27] Kahn and Taylor (1994).

1. Distort the high-capacity market, since other large users would presumably engage in efforts to obtain the same favourable rates;
2. Result in arbitrage, leading to lower retail prices, which would reduce the wholesale prices further if they are a set percentage of retail prices, and resulting in a kind of 'reverse squeeze' on the incumbent;
3. Reduce the incentive to improve efficiency if they lead to a reduction in retail prices which would also result in a reduction in wholesale prices; incumbents may then conclude that the best strategy would be to increase retail prices in order to increase wholesale prices.
4. Reduce incentives by entrants to construct physical facilities, if the wholesale discount is favourable to their competitors. It has been argued, on the other hand, that resale entry and the intensified competition it engenders are likely to enhance incentives for incumbents to maintain and upgrade their existing network facilities[28].

Thus, wholesale pricing creates a delicate 'knife-edge' problem. Set too low, the resale entrant cannot succeed. Set too high, resale becomes more attractive than physical entry. That situation also implies a continuance of a regulatory regime to set the discount since, in the absence of infrastructure competition, incumbent LECs would continue to control local distribution. The knife-edge dilemma may suggest a strategy of initially setting the discount large enough to encourage entry, but also at the same time establishing a fixed and definite term for its termination or phase-out, i.e., a sunset clause. The anticipation of the end of the regulated discount would then create incentives for entrants to construct alternative physical facilities. Another issue is that if the wholesale price were low, it would be then to the advantage of large users to avail themselves of that price. Hence, wholesale prices are generally made available only to telecommunications carriers, though the meaning of that term is not always clear. Universities provide telecom services to their students and faculty, and landlords furnish service to their tenants. To gain the advantages of interconnection prices hence requires some official recognition as a telecommunications provider.

While there is frequent agreement on the need to provide a discount to interconnecting wholesale services, the extent of the discount is usually disputed. Interconnectors and resellers advocate a substantial discount as the quickest method of encouraging competition. A discount that does not permit viable competition should therefore be presumed unreasonable. This brings them close to advocating that their survival (or even prosperity) is the test of a reasonable discount. Physical carriers, on the other hand, such as incumbent LECs, but also their infrastructure competitors such as cable TV companies, facility-based CLECs, and IXCs, favour low discounts since substantial discounts would provide disincentives for entrants to construct its own facilities and thus discourages

[28] Beard, et al. (1998).

facilities-based competition[29]. The measure for this wholesale discount has been 'avoided cost.' The ascertainment of such cost inevitably leads to regulatory and conceptual disputes. In the U.S., the FCC bootstrapped itself to state determinations and came up with a default range for the discount of 17–25 percent.

5.1.7. Efficient component pricing

An influential option for access pricing is the so-called *Efficient Component Pricing (ECP) Rule*. Under this option, aside from the direct incremental costs incurred by the incumbent in providing access, the entrant is required to compensate the incumbent for the loss of net revenue that its entry may cause. The interconnection price for an entrant's call would then be the average incremental cost, including all relevant incremental opportunity costs[30].

Advocates claim that ECP would improve or at least maintain current welfare through its four main properties:

1. Only interconnectors with lower incremental costs would be willing to enter the market.
2. New entrants would not affect the incumbent's revenues.
3. It does not interfere with cross-subsidisation (desired or otherwise).
4. It eliminates an entry barrier since incumbents would have no incentive to keep rivals out.

According to its proponents, ECP provides an approach that would closely parallel the methodology that a firm would employ in an environment when it sells its inputs to a firm that intends to compete in the final product market.

Criticism to this approach has been quite strong[31]. It was described, among others, as guaranteeing monopolists' profits against competitive losses; assuming zero-sum traffic; neglecting dynamic efficiency gains, and ignoring positive network externalities bestowed by entrants.

In response, Baumol and his co-authors agreed that, by itself, ECP would not result in a competitive pricing structure. In particular, initial charges must be constrained by market forces or regulation. Only if such condition was met would a 'level playing field' result and welfare gains be realised. ECP is not designed to do away with monopoly profits, and the problem lies in the incumbent being permitted to charge monopoly prices in the first place. Baumol assumes either a perfectly regulated market or 'a perfectly contestable market.'

Contestability means that competitors' could enter at will if the service was priced above competitive level by the incumbent. However, the competitors' ability to enter is precisely the issue in interconnection. If they could enter, Baumol's entire analysis is pointless, because market transactions would

[29] Kaserman and Mayo (1997).
[30] Baumol and Sidak (1994).
[31] See, for example, Economides and White (1995) and Tye and Lapuerta (1996).

determine interconnection prices. Interconnection pricing is a policy issue precisely because markets are not contestable.

In the real world, the weight of ECP was bolstered after a distinguished court upheld in 1994 Telecom New Zealand's opportunity-cost based charges. In the *Judgments of the Lords of the Judicial Committee of the Privy Council*, the Commonwealth's highest tribunal, separated two issues: monopoly rents, and abusive conduct. Even though it agreed partly with the New Zealand Court of Appeal on the matter of monopoly rents, its decision was that ECP could not be considered an abusive pricing principle:

> "Their Lordships are of the view that, apart from the risk of monopoly rents, the Baumol Willig Rule does provide a proper model for demonstrating what would be charged by the hypothetical supplier in a perfectly contestable market."

Yet for the U.K. itself, the regulatory agency OFTEL found several disadvantages with ECP. Especially where economies of scope exist for long distance and local services and where monopoly pricing is therefore likely, ECP would require an entrant to have its full costs lower than the incremental cost of the incumbent. OFTEL therefore concluded that ECP did not support the major objective of its interconnection framework, namely, to help the market move towards a more competitive environment.

In the United States, too, the FCC explicitly rejected the ECP framework with respect to interconnection and access to unbundled network elements, noting that the ECP methodology 'precludes the opportunity to obtain the advantage of a dynamically competitive marketplace'[32]. The FCC concluded that ECP does not replicate a competitive environment but rather (1) continues the inefficient and anti-competitive aspects of the existing price structure; (2) distorts competition by providing incentives for incumbent LECs to shift costs of competitive services to bottleneck services; and (3) preserves the status quo and serves as a barrier to entry.

5.1.8. Marginal cost pricing

In a fully competitive market, long-term marginal cost pricing is optimal in terms of economic efficiency. Marginal cost is defined as the cost of producing one more unit of output. The cost of increasing output by a given quantity is incremental cost, which approximates marginal cost if averaged across the increment. When applied to interconnection, incremental cost pricing will recover the additional costs the incumbent incurred by as a result of interconnection.

While the principle of long-run marginal cost pricing is simple, its conceptual and practical details are not. What is the long-term in a dynamic industry? Over what volume is it distributed? What are costs in a monopoly environment if a monopolist is inefficient? What is included in cost, given the frequent incentives to

[32] FCC (1996).

shift costs from competitive lines of business to price-regulated ones? What are the costs properly attributable to one line of business when some costs are 'joint and common'? Should marginal cost include the opportunity cost of foregone profits to the incumbent due to the interconnection?

A further question is whether to set long run incremental cost based on actual *historic* costs, present costs, or even future costs. Historic costs, also referred to as embedded costs, are the actual cost incurred to build the network. On the other hand, future *forward looking* cost methodologies do not involve the use of an embedded rate base, but rather postulates a hypothetical network based on near-term best-practice technology and efficient engineering. The regulatory search for the proper marginal cost led to long-run incremental cost (LRIC) methodologies. The U.S. variant is known as TELRIC, the European is as FL-LRAIC and the Japanese LRIC. For TELRIC, the cost basis is not the telecommunication *service* (for example, interstate access service) but rather the unbundled telecommunication *elements* that are used to provide the service (for example, for local loop or the local switch).

Incumbent LECs generally argue that future costs should be based on each incumbent's existing network technology, not on some idealised least-cost, most efficient network that may bear no relationship to existing operations. Furthermore, they argue that competitors would not invest in their own facilities if the alternative were an LEC element at a regulated price that is no higher than the least cost, most efficient provider. Entrants, on the other hand, prefer forward-looking prices at least as long as costs are declining over time. They contend that in a competitive market the most efficient provider will set the prices, regardless of such cost by the incumbent. Also, pricing based on the most efficient technology prevents cross-subsidies by the incumbent LECs, e.g., their charging the cost of facilities used to compete in other markets. In addition, entrants would be able to share in the economies of scale and scope of the incumbent, which limits the strategies available to the incumbent in preventing entry.

Forward-looking cost methodologies attracted criticism[33]. Setting prices equal to the forward looking cost will in many cases not allow the recovery of total costs, i.e., total revenues will fall short of total costs, especially with the transition from the previous system that tended to under-depreciate. Furthermore, even while average prices would approach cost, this does not necessarily force each individual element's price towards cost. There would still be varying price-cost margins among different customers and services. Incumbent LECs also argued that past investments, including inefficient ones, were frequently a direct result of regulatory requirements and that they therefore should be entitled to recovery.

In response, proponents of TELRIC argue that the price-cap recovery of embedded costs will lead to higher prices and result in the entrants over-building new capacities instead of maximising the use of existing facilities. To this,

[33] See, for example, Sidak and Spulber (1997).

incumbent LECs respond that the exclusion of embedded costs in the rate structure is confiscatory in nature and thus a 'taking' of property. TELRIC has also been criticised by proponents of 'real options' theory. Their basic critique has been that the FCC's pricing rules for unbundled network elements leads to prices that are too low. They contend that incumbents are required by regulation to give a free option to entrants, the option being the right but not the obligation to obtain the unbundled element[34].

5.2. Arbitrage

One can set interconnection prices administratively or by market forces. The former is intrusive, while the latter requires competition and the absence of market power. Where competition is partial, its role can be extended through arbitrage and benchmarking, such as 'Third Party Neutrality.' (TPN). As long as some local network segments are competitive, the access charge system can be simplified. Networks can charge any price and select their customers subject only to the following principles:

1. Networks cannot prevent arbitrage by discriminating against their customers' customers.
2. The prices for monopolistic network segments where competition exists is based on those prevailing for similar competitive segments (i.e., on benchmarks).

These are simple rules. The second would wither away with competition. The first would join various similar rules that exist in commercial transactions, such as the holder-in-due-course doctrine, whose purpose is to facilitate transactions in the economy.

5.3. Incremental cost

It is one thing to speak in the abstract of total cost, historic cost, average cost, and so on. It is quite another matter to define and measure them. In a functioning market, cost need not be demonstrated, only acted upon. However, when, as is more likely, an interconnection charge is either regulated or its feasibility is subject to appeal before a regulatory agency, cost needs to be defined and justified.

Incumbent network operators have strong incentives to assign high portions of joint costs to interconnection services. The problems for the regulator are therefore to identify which costs are relevant to interconnection, how large they are, and how to allocate them. Economically efficient cost allocation would suggest that the costs of services should be borne by those who cause them. But as applied

[34] Hausman (2000).

to the pricing of interconnection, it can be difficult to operationalise this principle. This is due to conceptual questions – which costs should properly be included? – as well as for factual reasons – what are the figures, and are they reasonable? The problem is further complicated insofar as companies are asked to disclose proprietary cost information that competitors may find useful.

The tracking of cost information is difficult. The integrated structure of carriers, together with the complexities of joint and common cost allocation, make it hard to determine cross-subsidies and cost-shifting. This is compounded where equipment, facilities and manpower is shared by the regulated as well as the unregulated operations. In the United States, the FCC and the state commissions therefore implemented a complex and regulated set of structural and non-structural accounting rules that try to ensure a proper allocation of costs and revenues. This is the *Uniform System of Accounts*. All major local telephone companies have to comply with the USOA's basic accounting rules. Accurate records must be kept for revenues, operating costs, depreciation expenses, and investment in plant and equipment. Also, a distinction must be made between regulated and non-regulated businesses.

Costs have been divided into traffic-sensitive (TS) and non-traffic-sensitive (NTS) costs. The fixed NTS costs comprise, among others, the local line, maintenance costs and a portion of local exchange switching equipment. The variable TS costs cover switching and trunking plant and equipment that is shared by all users. The methodology distinguishes among separate variable cost elements, such as line termination functions in the LEC's end office, traffic sensitive switching equipment, directory assistance, inter-exchange facilities, and common transport facilities. The large telephone companies then provide this data on the FCC's Automatic Reporting Management Information System (ARMIS).

As if this is not enough, an additional critical issue exists in the United States, in the 'jurisdictional separation' of inter- and intrastate costs and revenues[35]. Costs assigned to the interstate jurisdiction are recovered from interstate service such as access charges, (i.e., the connection between local and long distance services). But what is that interstate share, given the integrated nature of a communications network and its use for both long-distance and local calls? After much dispute, the overall interstate share of the larger LECs' switched network costs was fixed, by a compromise that was based on politics rather than economics or technology, at 25 percent, even though total interstate switched traffic was only 14 percent of all switched traffic.

A second major approach to determining cost is to skip actual cost measurements for generalised ones. Instead of determining what cost had been, one calculates what it should be, looking forward. These numbers can either derive from armchair analysis, but more likely originate from 'proxy' engineering models. These models, initially designed in the United States to determine the appropriate

[35] NARUC (1971).

interstate universal service burdens and allocations generally use geographically based units to determine the standard costs of serving customers within a set area. Not only must the area be representative in terms of terrain and customer dispersion, but it must also approximate the size in which engineers would be comfortable in making provisions for equipment. For example, one of the proxy models in the United States, the Benchmark Cost Model (BCM), that was supported by several ILECs, looked for the average costs required to serve residential customers within census block groups generally containing 250 to 550 households which were assumed to be evenly distributed.

Various approaches have various financial assumptions. For example, the HAI Model, supported by several interconnectors, depreciates equipment over 18 years, with the cost of equity at 11.25 percent, whereas the Cost Proxy Model (CPM) supported by other ILECs uses a 12-year cost model depreciable life.

The FCC developed a hybrid proxy based on the outputs of several models proposed to it, along with a local loop design and clustering algorithm developed by its own staff. This hybrid proxy cost model was adopted in 1998[36].

The proxy cost models have also been used, with modifications, to determine the cost of the unbundled network elements (UNEs), a task delegated by the FCC to the state utility commissions, under the condition of utilising a forward-looking cost methodology, an economics-based rather than regulatory depreciation, and an risk-appropriate cost of capital. The state-determined UNE cost over a wide range (10:1), suggesting that any methodology can be fitted to the policy goals. A major source of differentiation was the treatment of 'non-recurring cost,' which incumbents had incentives to inflate. Engineering-based cost models, despite their scientific claims, are only as credible as their underlying assumptions. Supporters of the different proxy models advocate a model that favours their own interests. In the process of balancing pressures and adding 'realism,' models have lost transparency to users and policy makers. They acquired the unmistakable flavour of an administrative tool of a planned economy, even if they were adopted in pursuit of market competition.

How does all this add up? An analysis such as Mark Armstrong's companion piece in this volume, can throw light on the consequences of several of these approaches. But in another sense, the vigorous discussion over pricing principles also shows that while ideas matter to the world, the world matters even more to the visibility of ideas. In telecommunications, these ideas materially affect interconnection charges paid by some companies to others. For some American long-distance companies, these payments used to account for about 40 percent of their overall expenditures; for the local exchange companies, the receipts were over 20 percent of their revenues. The pricing and costing principles also relate directly to the payments that various companies make towards the financing of the universal service. The magnitude of that redistributive system has been estimated, de-

[36] FCC (1998), Sharkey et al. (1992), Sharkey (2001).

pending on definition, methodology, and interest, to be anywhere between about $4 and $20 billion. In many cases, the setting of the interconnection charge is therefore a matter of corporate life and death. Given those stakes, it is not surprising that supportive ideas are in demand by each side and that they receive a wide play by their proponents.

Different economic models lead to different conclusions. The efficient component-pricing (ECP) rule has been advocated by several distinguished economists. It is advantageous to the incumbent local exchange companies charging high prices. Other pricing models result in low interconnection prices, and are therefore favoured by new entrants. Forward-looking long-run incremental cost (LRIC) is such an approach, and it, too, is supported by equally distinguished scholars. It is supplemented by a planned-economy style, engineering-based proxy cost models that are advanced by the staunchest advocates of free markets. Various experts are lining up before the regulatory decision-makers, brandishing competing theories with well-compensated passion. Who is right? Perhaps a better question is what the policy goal is.

Thus, when the policy goal is to expand basic telephone service or to keep basic telephone prices low, regulators will be supportive of the incumbents as long as these recycle their gains into wide and affordable connectivity. In that situation, the cost models picked will tend to be along the lines of efficient component pricing or of distributed cost pricing. Where large customers are to be favoured, Ramsey pricing provides an efficiency rationale. With vigorous local competition as the goal, regulators have adopted marginal cost models whose fundamental advantage to entrants is that they are lower in price by reducing or postponing their contribution to fixed costs. And when regulators have tried to further accelerate the pace of entry into local competition, they extended this approach into cost that is forward-looking. There are some good theoretical arguments for such a methodology, but it is doubtful that this approach would have been chosen if the price would not trend conveniently down, but was instead going up and thereby slowing down entry.

It would be easy to conclude that the carriers of ideas are merely the champions of the various carriers of transmission. Yet, if we measure new concepts only by the yardstick of *cue bono*, debates over ideas would be pointless. Out of thesis and antithesis, however motivated, a higher form of understanding will hopefully emerge.

6. Interconnection around the world

A wide range of approaches to interconnection exists within different countries and among them. Yet most issues are similar and inclusive: the impact of emerging network competition; the pricing by incumbent operators; the determination of the true cost of the underlying infrastructure; the assurance of

transparency in incumbent operation; the striking of a balance between general government regulation and competitors freedom to negotiate their own deals; the sharing of the burden of universal service obligations; the symmetry of payment obligations; the promotion of entrants in order to protect competition, yet without providing them with guarantees of survival. While the basic issues are common, different countries are moving at varying speeds and with different emphasis on competing policy goals. This is based on history, politics, economics, the state of the evolution of the public network, and the development of an information-based economy. Several countries have taken the lead in pushing the envelope of activist interconnection policy. They tend to be those countries with the longest experience in telecommunications liberalisation.

In the process of moving ahead at different rates of change, divergent policies are creating international tensions. The interconnection of national networks with each other, an unusually harmonious sector for international co-operation since the middle of the 19th Century, has become a source of struggle. For a long time, international institutions provided the mechanism for harmonising restrictive cartel behaviour in interconnection. But no longer.

The International Telecommunication Union was created, to a significant extent, to manage early international interconnection arrangements. For many decades it was supportive of a system of an international cartel of national monopolies. It firmly opposed, for example, the interconnection of leased lines with public networks, as well as the resale and sharing of leased line capacity. In the 1990s however, the ITU process became more accepting of privatisation, liberalisation and competition.

Similarly, Intelsat, the major international satellite consortium, had embodied the collaborative international system, with all of its strengths and weaknesses. For a long time Intelsat had legal exclusivity over international civilian satellite communications. Intelsat was used to operate a global gatekeeper system. Interconnection into Intelsat infrastructure was granted only to a single entity per country, usually the monopoly incumbent operators. No other carrier could interconnect directly into Intelsat, and only official carriers could interconnect into the designated carriers. The notion of a large user uplinking directly to Intelsat was sheer heresy. The purpose of this arrangement was to tightly control international satellite traffic, especially in its economic dimensions. International calls were enormously profitable, but their prices could not be maintained in competition. This indeed happened in their 1990s when Intelsat's hold over international traffic collapsed. In consequence, Intelsat shifted its role and became a market participant rather than cartel enforcer.

In Europe, the EU Commission pursued two goals, sometimes at contention with each other: the *harmonisation* of European rules of telecommunications; and the *liberalisation* of telecom markets. Liberalisation provided the banner of substance but harmonisation, under control of Brussels, was also the bureaucratic

agenda. The principles for harmonisation in domestic interconnection was laid out in the 1990 ONP (Open Network Provision) Framework Directive (1990), and fleshed out in the subsequent ONP voice telephony directive. Interconnection and its pricing were the most controversial issues in that directive. The traditional monopoly PTOs demanded contribution for their infrastructure investments through recovery of historic cost, while the inter-connectors advocated a non-discriminatory and transparent 'equivalent interconnection.' The EU Commission fudged this, giving leeway to national governments for the actual details of interconnection, access charges, and dispute arbitration. The Commission's role was therefore to regulate regulators rather than market participants[37].

Unfortunately, the ONP voice telephony directive had the dubious distinction of being the very first piece of legislation rejected by the European Parliament under its newly acquired powers of co-decision created by the Maastricht Treaty. The directive was rejected by the European Parliament, which found a lack of consumer protection provisions and weak transparency requirements. Soon thereafter, the European Parliament adopted an updated proposal (1995).

Telecom companies which account for over 25 percent of national markets are presumed to have sufficient market power (SMP) and must fulfil certain obligations. These include meeting all reasonable requests for access to their networks, non-discrimination, unbundling, publication of a reference interconnection offer (including price lists), cost oriented tariffs, and a transparent accounting systems, including accounting separation in some cases. Such accounting separations require operators to implement appropriate cost allocation methodologies, with a preference for long-run cost methodologies. National regulators can intervene in disputes and inspect any agreement.

The European Commission also adopted a recommendation on interconnection pricing which supports the eventual use of a forward-looking long-run average incremental cost model (FL-LRAIC), with 'best current practice' charges in the interim.

Once Europe, the United States, Japan, and several other major countries had moved to a system of an interconnected network of networks, others followed. In 1997 the member countries of the World Trade Organisation (WTO) concluded a multilateral deal aimed at liberating international trade in basic telecommunications services, with direct impact on interconnection. The deal (technically not an 'agreement') comprised of 55 schedules, covering 69 governments, of legally binding commitments to open their basic telecommunications markets to foreign competition. The national schedules are an integral part of a larger treaty, the General Agreement on Trade in Services (GATS).

Telecommunications services were one of the first sectors negotiated under the new WTO regime. A major policy breakthrough in the process was an agreement to include an American-authored reference paper on regulatory reform. The paper

[37] Austin (1994).

had direct relevance to interconnection issues. It laid out several key principles. Incumbent network operators must provide market entrants with interconnection at any technically feasible point in the network. Such interconnection must be provided under non-discriminatory terms, conditions and rates, and should be of a quality no less favourable that the provider gives its own services. Moreover, interconnection rates must be cost-oriented, transparent, and where economically feasible, unbundled. A dispute mechanism administered by an independent body was also called for. The agreement was significant because it institutionalised for the first time a system of multilateral mutual surveillance and a framework for bilateral bargaining over market entry. But, reality may disappoint. There is indeed an agreed-upon time schedule, but it applies only to the WTO segment of any dispute. The WTO is an organisation of governments, not a civil court. A firm must convince its national governments to back its claim and present it to the other governments and the WTO arbitration panel.

In the past the model of international institutional co-operation aimed to protect national monopoly systems, and was exemplified by the 'old' ITU and Intelsat. Their interconnection policies were squarely aimed to preclude competitive entry. But by the beginning of the new century, the international institutions of telecommunications had moved to a liberalising role, with more open interconnection arrangements at its centre. In these bodies, interconnection policy had proceeded to its second stage, that of opening national and international markets.

7. Interconnection and common carriage

What impact does interconnection of networks have on content? Would information flow with greater ease or with more restrictions? Two types of legal status apply to various carriers. The *private carriage* approach is one of contract and property. The private owner of an electronic conduit can utilise its capacity as it sees fit, transacting with those partners it wishes to engage with and carrying only the information content it wishes to accept. In contrast, *common carriage* has been the underlying system for the flow of information over the telephone network. It reduced the ability of network providers to discriminate among similar customers, and to select their customers[38]. It required a carrier to accept any lawful content.

Precursors to common carriage in telecommunications go back to the Roman Empire and the legal obligations of ship owners, innkeepers and stable keepers. In England, the early common law placed certain duties on mostly infrastructure businesses which were considered 'public callings.' 'Common' in that context meant 'open to serving the general public' or 'general.'

[38] Noam (1994b).

For centuries, common carriage principles have played an important role in the infrastructure services of transportation and communications. This system was created to assure that all customers seeking an essential infrastructure service and willing to pay the established price would not be denied lawful use of the service nor otherwise be discriminated against. For one hundred years these principles have facilitated telecommunications users' access to the public networks. Carriers were severely limited in selecting their customers and their usage of the network, including content types as long as they are lawful. A private telecommunications carrier obtained certain benefits, including limited liability for the consequences of its own actions. It also often received, by statute, powers of eminent domain, use of public rights-of-way, and even protection against competition. Today, as networks interconnect physically, functionally, and financially, they must also converge in terms of the rules under which content flows over them. While the old system of segregated networks made it possible to segregate content flow rules, this is becoming difficult as network interconnect. Starting in the 1980s, telephone common carriers, reacting primarily to outside pressures and concerns about their corporate image, attempted to ban or restrict sex lines based on content, even where the content of the messages was legal. One important aspect of common carriage is that it facilitates interconnection and hence competition. This is because under common carriage, access is provided to all customers, even where they are the competitors of the carrier. This reduces the entry barriers for competitors, since they can utilise elements of the common carrier. For example, under this principle, MCI was able to reach its customers over AT&T's local networks rather than having to first build its own local distribution facilities. Hence, common carriage reduces monopoly power, though its availability may also reduce the incentives to competitive entry or to upgrades.

With the advent of interconnectivity the demise of common carriage becomes a distinct possibility. There are several reasons a common carrier cannot use differentiated pricing in the same way that a private contract carrier can. In particular, a private contract carrier can pick customers and prevent arbitrary. If interconnection links these two legal regimes, what will be the impact? In head-to-head competition, the restrictions on common carriers disadvantage them against private carriers, all other things equal. Common carriers qua carriers will not become extinct, but they will increasingly conduct their business as private carriers, and common carriage as such will disappear over time. This will not happen overnight, of course. But the basic dynamics will sooner or later assert themselves.

A common carrier must serve a private contract carrier, but not vice-versa. These conceptual discussions have a very practical dimension, which revealed itself as cable TV companies moved beyond their traditional activities. In 1999, when AT&T bought the cable TV giant TCI, it also acquired TCI's @Home, a cable-based ISP and on-line content provider using high-speed cable modem access to @Home's proprietary content as well as to the Internet. The ISP industry, led by AOL protested the arrangement. The problem was that TCI

bundled its cable service with @Home, and cable subscribers who wanted to access other ISP and content services from another provider would have to pay an extra fee, making them uncompetitive with @Home. The rival ISPs therefore asked the FCC to require an equal access arrangement from AT&T/TCI for any ISP that desired such access. In effect (though not in these words), they asked the FCC to impose a quasi-common carrier status on cable carriers, or at least on their packet transmission. Such a policy would inevitably also apply to cable-based telephony. The ISP problem was upheld by several lower courts. An appellate court overturned in 2000. AT&T won on jurisdictional grounds, but it could hardly been happy about its victory, because the court classified Internet access as a 'telecommunications service' subject to federal regulation, which meant that it was potentially subject to all the regulations aimed at telephone companies, such as unbundling and access requirements.

The demise of common carriage raises the question of impaction on the free flow of information when any carrier in a chain of transmission could impose its own standard of acceptable content. One alternative approach to the regulatory requirement of Common Carriage is arbitrage, 'Third party neutrality' (TPN), as a substitute for common carriage. It imposes no common carriage obligation on any carrier. However, if a carrier interconnects another carrier or user, it cannot restrict them from contracting with third parties. The carriers cannot discriminate against its customers' customers. This, in effect, creates arbitrage in content access, and makes it difficult to discriminate in price or through selective interconnection. Third-party neutrality thus ensures a non-discriminatory flow of information in an environment where carriers can contract freely, as long as there is at least some access node that is not controlled by a carrier.

8. The future of regulation of interconnection

Our discussion has taken us from *history* – where we identified control over interconnection as the key lever in the traditional establishment of monopoly; to *policy*, with interconnection as the battleground for creating and blocking competition; to *international relations*, where interconnection has become a global issue; to *economics*, where the pricing and costing of interconnection are matters of survival.

The remaining question is whether in a fully competitive environment, any residual interconnection rules are needed, such as requirements to interconnect or to avoid discrimination.

The issue is not normative but positive. Regulation exists not because of bureaucrats who cannot let go. It is created largely as a political response to interest groups. These interest groups will not disappear and new ones will emerge. In a democratic political system a majority coalition always wants something from the minority. In telecommunications, this has been the underpinning of such policies as universal service, rate averaging, high-cost area

redistribution, and so forth[39]. With telecommunications becoming ever more important, not having full connectivity to the new and powerful means of communication becomes a major disadvantage. That is why, inevitably, the definition of universal service will expand, even as competition makes basic services more efficient. The introduction in the U.S., of a favourable 'e-rate' for Internet access by schools, libraries, and hospitals, is an early example. Other issues are consumer protection, privacy, or rural service.

Throughout this survey, we have described the changing use of interconnection. In its first stage, it was *pro- incumbent*, and used to establish monopoly control. Much later, the goal was *pro-competition*, and interconnection was used to pry open the network environment to new entrants. When local competition was slow in letting entrants to catch up with the entrenched incumbent, interconnection policy moved to its third stage, in which it aimed at *market-control*.

In most countries, even after a number of years of competitive entry, the incumbent still is dominant in most traditional market segments. Supportive interconnection policies can accelerate competition. Yet it becomes a problem if they are open-ended and without explicit phase-out. In that case, it will be difficult to terminate the rules in the future because entire clusters of firms will be dependent on them for survival. Even if some of the entrants became vigorous enough to be able to compete without regulatory help and to negotiate non-coercive interconnection agreements, there will always be competitors on the margin who would not survive the abolition of supportive interconnection rules. Their impending demise would then create desperate fights to maintain the rules, and to public arguments that their abolition would reduce competition by reducing the number of competitors.

Hence, it becomes more difficult to exit interconnection regulation than to enter it, unless the exit scenario is spelled out clearly in advance. This provides incentives to the entrants to plan for a competitive scenario where they are on their own. Within reason, such clarity is much more important than the length of the transition.

Entrants do not like specific exit scenarios because they fear that it permits incumbents to engage in dilatory practices. But that is only true if incumbents can delay with impunity and without cost. Hence, the exit timing needs to be tied to equally specific schedule for the incumbent to fulfil a set of requirements in the provision of the elements of actual interconnection, such as support systems, number portability, etc.

If interconnection rules are temporary only, there is less need for them to be overly complex. In the United States, the battles over the precise nature of interconnection methodology and its prices models have delayed actual local entry, and will continue to do so in the future as the details of the rules will require further

[39] Noam (1994c).

clarification, proceedings, regulation, litigation, and legislation. A temporary solution can, on the other hand, then be structured as a relatively simple system.

The alternative - interconnection regulation into the distant future without a clear exit scenario - maintains the regulator in a position to establish entitlements, to create client relations by the industry, and to maintain its role of perpetually indispensable arbiter. This is not a scenario of deregulation and market competition, but of a micro-managed telecommunications market and industry dependency into the distant future.

Furthermore, it is likely that the significance of government's wielding the tool of interconnection policy will increase rather than decline. Even if telecommunications markets become more competitive, there will be no shortage of problems beyond bottleneck power, the present main rationale. Consumer protection is one example. The affordability of certain socially valuable services is another. To deal with those issues, government requires a policy instrument, and the control over interconnection provides it. To add to its usefulness, it is largely 'off-budget,' requiring no outlays, only rules and mandates. With those advantages, it is difficult to imagine that interconnection will be withdrawn from the regulatory arsenal. If anything, its role will grow.

References

Austin, M. T., 1994, Europe's ONP bargain: What's in it for the users? Telecommunications Policy, 18, 97–113.

Armstrong, M., 1998, Network interconnection in telecommunications, Economic Journal, 108, 545–564.

Armstrong, M., C. Doyle and J. Vickers, 1996, The access pricing problems: A synthesis. Journal of Industrial Economics. 41, 335–360.

Armstrong, M., 2002, The theory of access pricing and interconnection, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science B. V., 295–384.

Baake, P. and T. Wichmann, 1999, On the economics of Internet peering, Netnomics, 1, 89–105.

Baumol, W. J. and D. F. Bradford, 1970, Optimal departures from marginal cost pricing, American Economic Review, 60, 265–283.

Baumol, W. J. and J. G. Sidak, 1994, The pricing of inputs sold to competitors, Yale Journal on Regulation, 11, 171–202.

Beard, T. R., D. Kaserman and J. Mayo, 1998, The role of resale entry in promoting local exchange competition, Telecommunications Policy, 22, 315–326.

Braeutigam, R., 1980, An analysis of fully distributed cost pricing in regulated industries, Bell Journal of Economics, 11, 182–196.

Brock, G. W., 1993, The U.S. telecommunication policy process, A study in decentralised public decision making, mimeo.

Cave, M., 1994, Regulating service provision in mobile telecommunications: Some lessons from British experience, mimeo.

Economides, N. and L. J. White, 1995, Access and interconnection pricing: How efficient is the efficient component pricing rule? The Antitrust Bulletin, 16, 271–284.

Federal Communications Commission, 1997a, Access charge reform, Docket 97–158.

Federal Communications Commission, 1997b, Access charge reform, Docket No. 96–262, No. 96–45.
Federal Communications Commission, 1982, Exchange network facilities, Docket No. 78–371, 90 F C.C.2d.
Federal Communications Commission, 1967, in Re microwave communications, Inc. Docket No. 16509, 64 F C.C. 2d 979.
Gabel, D. and D. F. Weiman, 1994, Historical perspectives on interconnection between competing local exchange companies, Columbia Institute for Tele-Information, Working Paper.
Hausman, J., 1999, The effect of sunk costs in telecommunications regulation, in J. Alleman and E. Noam, eds., The New Investment Theory of Real Options and its Implications for Telecommunications Economics, Boston: Kluwer Academic Publishers.
Huber, P. W., 1994, Competition and open access in the telecommunications market of California, Unpublished report.
International Telecommunication Union, 1995, The changing role of government in an era of telecom deregulation. Interconnection: Regulatory Issues, Geneva, Switzerland: International Telecommunication Union.
Kahn, A. E., 1988, The economics of regulation: Principles and institutions, Cambridge, MA: MIT Press.
Kahn, A. E. and W. E. Taylor, 1994, The pricing of inputs sold to competitors: A comment, Yale Journal on Regulation,11, 225–227.
Kaserman, D. and J. Mayo, 1997, An efficient avoided cost pricing rule for resale of local exchange telephone services, Journal of Regulatory Economics, 11, 91–107.
Laffont, J.-J. and J. Tirole, 1994, Access pricing and competition, European Economic Review, 38, 1673–1710.
Laffont, J.-J. and J. Tirole, 1996, Creating competition through interconnection: Theory and practice, Journal of Regulatory Economics, 10, 129–145.
Laffont, J.-J. and J. Tirole, 2000, Competition in telecommunications, Cambridge, MA: MIT Press.
Lehr, W. H. and M. B. H. Weiss, 1996, The political economy of congestion charges and settlements in packet networks, Telecommunications Policy, 20, 219–231.
Mueller, M., 1988, Open interconnection and the economics of networks: An analysis and critique of current policy, Columbia Institute for Tele-Information Working Paper Series.
National Association of Regulatory Utility Commissioners, 1971, Separations, Manual-Standard Procedures for Separating Telephone Property Costs, Revenues, Expenses, Taxes, and Reserves, Washington, D.C.
Noam, E., 1994a, Beyond liberalisation: From the network of networks to the system of systems, Telecommunications Policy, 18 (4), 286–294.
Noam, E., 1994b, Beyond liberalisation II: The impending doom of common carriage, Telecommunications Policy, 18, 435–452.
Noam, E., 1994c, Beyond liberalisation III: Reforming universal service, Telecommunications Policy, 18, 687–704.
Noam, E., 2001, Interconnecting the network of networks, Cambridge, MA: MIT Press.
Sharkey, W. W., 2002, Representation of technology and production, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science B. V., 179–222.
Sidak, J. G. and D. F. Spulber, 1997, Giving, takings, and the fallacy of forward-looking costs, New York University Law Review, 5, 1068–1164.
Tye, W. and C. Lapuerta, 1996, The economics of pricing networks interconnection: Theory application to the market for telecommunications in New Zealand, Yale Journal on Regulation, 13, 419–500.
Vogelsang, I. and B. M. Mitchell, 1997, Telecommunications competition: The last ten miles, Washington, DC: AEI Press.

Weare, C., 1996, Interconnections: A contractual analysis of the regulation of bottleneck telephone monopolies, Industrial and Corporate Change, 5, 963–992.
Willig, R. D., 1979, The theory of network access pricing, in H. M. Trebing, ed., Issues in Public Utility Regulation, Proceedings of the Institute of Public Utilities Tenth Annual Conference, East Lansing, MI: Michigan State University Press.
Wright, J., 1999, International telecommunications, settlement rates and the FCC, Journal of Regulatory Economics, 9, 71–92.

This Page Intentionally Left Blank

Chapter 10

UNIVERSAL RESIDENTIAL TELEPHONE SERVICE

MICHAEL H. RIORDAN*

Columbia University

Contents

* The author acknowledges helpful comments from seminar participants at the University of Pennsylvania and Columbia University, and from Martin Cave, Robert Crandall, Patrick DeGraba, Gerald Faulhaber, Chris Garbacz, Jerry Hausman, David Kaserman, Lester Taylor, Peter Temin, and Ingo Vogelsang, while retaining full responsibilility for the final product. Saurabh Jain, Wenying Jiangli, Seth Seabury, Govert Vroom, and Pearl Wang provided useful research assistance for different parts and stages of the project.

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

Universal service is a chameleon-like phrase. It refers generally to widespread access to and affordability of telecommunications services, but it takes on different meanings depending on the time and the place, and the particular policy debate. AT&T president Theodore Vail coined the phrase in 1907 to refer to the company's goal of achieving an integrated centrally-controlled telephone network, but today in the United States and other developed countries the phrase essentially means high household telephone penetration (Mueller 1997). In less developed countries, where telephone penetration is low, the phrase more likely means good access to pay telephones (Hudson 1995). Recent universal service initiatives in the United States subsidize high-speed Internet access for schools, libraries, and health centers (Hausman 1998). And in the blue sky of the future, universal service may come to mean high residential penetration of broadband Internet access.

Since this landscape is too big to cover succinctly, this chapter focuses on the "paradigm problem" of advancing and maintaining universal service for basic residential telephone services in the United States in the late 20th century. The focus seems appropriate, if for no other reason than because this is where academic economic research has concentrated its attention. Moreover, some of the issues addressed by the chapter have wider applicability. For example, there is a "deadweight loss" of economic efficiency from taxing regular telephone service in order to subsidize advanced services (Hausman 1998). The chapter makes some international comparisons, and mentions a few emerging issues, but the reader is forewarned not to expect too much on these fronts.

Universal residential telephone service is an important and complex policy issue because large amounts of consumer welfare and corporate profits are at stake in the design of regulatory policies in the pursuit of universal service (Hausman 1998), and because important noneconomic values, like political democracy and social cohesion, are prominent in the policy debates. This volatile mix of elements makes for highly charged political debates on universal service policies, often with the Federal Communications Commission (FCC) at the center[2]. Economic arguments matter in these debates, even when noneconomic values have great salience, making universal service a worthy policy problem for applied economic analysis. What are the economic determinants of telephone penetration? What are the economic arguments for and against universal service policies? What is the most efficient way to achieve universal service goals? How successful are actual universal policies at increasing telephone penetration? The purpose of this chapter is to assess the current state of economic knowledge about universal service, and

[2] The FCC has various policies designed to promote universal service: subsidies for schools, libraries and rural health centers; support to carriers serving high cost areas; subsidies for low income consumers. See http://www. fcc. gov/ccb/universal_service/

to point out needs for further research. The chapter mainly restricts its attention to published economic research which presumably has been vetted by some form of peer review.

Section 254 of the 1996 Telecommunication Act directs the FCC and the states to adopt policies "for the preservation and advancement of universal service . . ." and defines universal service as "an evolving level of telecommunications services that the Commission shall establish periodically . . ." So far, the FCC has defined universal service essentially to encompass basic residential telephone services (Federal Communications Commission 2000). The language of the Act suggests that universal basic telephone service has been substantially but perhaps incompletely achieved in the United States. Figure 1 confirms this idea by showing that household telephone penetration has remained over 90% for more than a quarter century, and today approaches 95%[3].

Behind this rosy aggregate picture, however, there is considerable regional and local variation. The map in Figure 2 shows that penetration rates varied significantly across the states in 1990, ranging from 87.4% in Mississippi to 97.9% in Maine[4]. The variance is even greater at the county level, where penetration ranges from 40.3% in Apache County, Arizona to 99.5% in Waukesha County, Wisconsin. Mueller and Schement (1996) find large variations in penetration rates among neighborhoods of a single city. At the census block level, penetration varies between zero and one hundred percent.

The United States has one of the highest household telephone penetration rates in the world. Still, some other developed countries enjoy a higher aggregate household penetration rate, e.g. Canada has maintained penetration over 98% through the 1990's. Moreover, while household telephone penetration has remained relatively flat in the U.S. in the 1990s, it has increased significantly elsewhere, e.g. in France, from 94% in 1990 to 98% in 1997 (International Telecommunications Union 1999). Thus, it appears that more could be done to advance universal residential telephone service in the United States. Questions for economists are "How and at what cost?" and "Do the benefits outweigh the costs?"

The universal service problem for basic residential services has several dimensions, and the balance of the chapter is organized accordingly. Section II presents empirical evidence on the determinants of telephone penetration rates in the U.S. in 1990. The analysis shows that most of the variation in telephone penetration in the United States is explained by demography and climate. Cost proxies explain a statistically significant but quantitatively small fraction of the variation in

[3] This chart is constructed from various Census Bureau and FCC data sources, and contains linear approximations for some years to deal with missing and inconsistent data. Details of the construction are available are from author upon request. See FCC, "Trends in Telephone Service," March 2000, Wash D.C. for a discussion of subscribership data.

[4] This is based on 1990 census data. See Dyer (1997) on regional variation in penetration rates in the United Kingdom.

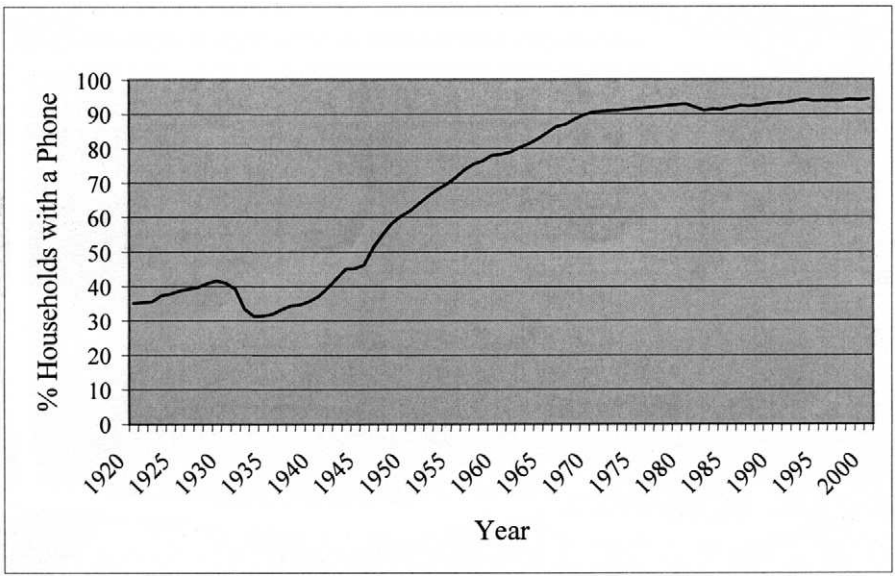


Fig. 1. Telephone penetration in the United States, 1920–2000.

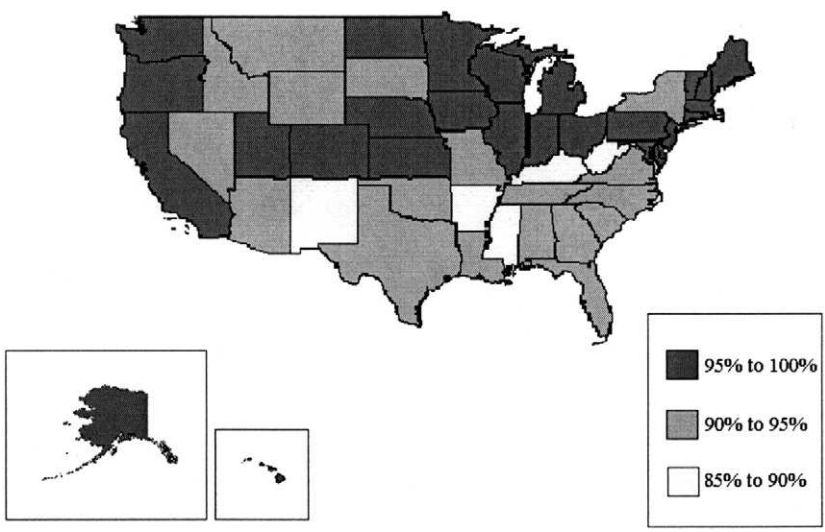


Fig. 2. Telephone penetration by state, 1990.

penetration, and there is some slight evidence of local network externalities boosting penetration. While there remain significant differences between the states even after controlling for these factors, it appears that superior state regulatory policies can explain at most only a few percentage points of universal service performance. Section 3 reviews the normative economic theory of telecommunications pricing and its implications for universal service. Scale economies and especially network externalities provide theoretical rationales for departures from strict cost-based pricing, even though such departures sacrifice economic efficiency on some margins. Economic theory also demonstrates that optional service plans and low-income and high-cost universal service support potentially are valid methods of price discrimination in the pursuit of universal service goals. Section 4 reviews published empirical evidence on the performance of actual universal service policies. This limited evidence shows that low-income and high-cost subsidy policies are at best only marginally effective at advancing universal service. Section 5 summarizes and draws conclusions.

2. Telephone penetration in the United States

Although approximately 95% of American households have a telephone, penetration varies significantly from place to place. Figure 2 illustrates different penetration rates in different states. Is this variation due to differences in population characteristics and other factors affecting the demand for telephone service, or differences in costs and regulatory policies affecting the price and availability of service? This section explores this question with a reduced form regression analysis. The purpose of the analysis is to identify and interpret some stylized facts, and to motivate the possibility that differences in state regulatory policies matter for the achievement of universal service goals.

Schement (1995) uses FCC and census data to describe the characteristics of households lacking telephones. The data show that the achievement of universal service varies across population groups. For example, the poor are less likely to have a telephone, as are blacks and Hispanics[5]. This kind of descriptive analysis is suggestive, but could be misleading, and leaves open important questions. For example: Are black households less likely to have a telephone because of different tastes, or because blacks tend to have lower incomes and telephone service is a normal good, or because blacks are discriminated against in the provision of telephone service? Or do blacks tend to live in states with less aggressive policies for promoting universal telephone service? Regression analysis is the appropriate tool for disentangling these effects.

[5] Schement, Belinfante and Povich (1997) provide a more detailed analysis showing among other things that households receiving various forms of public assistance have lower penetration rates.

A priori it seems plausible that demography might explain much of the geographic variation in penetration rates. Column (I) in Table 1 reports a regression equation explaining the telephone penetration rates of 1990 census block groups (CBGs) as a function of selected population demographics[6]. The definitions of variables and summary statistics are in an appendix. The numbers in parentheses are t-statistics, with the usual interpretation that a t-statistic above approximately 2.0 indicates statistical significance above a 95% confidence level. At given prices, shifts in the demographic composition of a group of consumers can be expected to shift the community demand for telephone service and change the penetration of service within the community. Nevertheless, the regression must be interpreted cautiously because it does not control for prices. The regression equation can be interpreted as capturing the pure effect of demand shifts on telephone penetration only if demand is price inelastic or if price differences are uncorrelated with population demographics. The results are broadly consistent with demand studies of penetration that do control directly for prices (Crandall and Waverman 2000; Taylor 1994; 2000).

Two things about the regression are striking. First, as expected from Schement's descriptive analysis, poverty is a major predictor of low CBG penetration. An income redistribution that would lower the poverty rate of a CBG by one percentage point, while holding its median income constant, would add 1/4 percentage point to telephone penetration. FCC Lifeline and LinkUp policies, discussed later in more detail, are designed to make telephone service more affordable to low income households. Second, Native American populations have much lower telephone penetration than other population groups, even after controlling for poverty, median income, education, and other demographics. It is not clear why this is the case. Do Native Americans place less value on telephone service, are they victims of discrimination, or is service more expensive or less available in areas occupied by Native Americans? Recently, the FCC targeted increased subsidies at federally-recognized Indian tribes, on grounds that the 47% average telephone penetration for this consumer group is partly due to expensive and unavailable service[7,8].

Other demographic characteristics of CBG populations influence penetration noticeably but less dramatically. The estimated effects are generally consistent with published descriptive analyses (Schement 1995; Schement, Belinfante, and

[6] All of the data for this regression equation are from the 1990 census. This is a weighted least squares regression which adjusts for the varying population sizes of CBGs. For a description of this procedure see Greene (1993).

[7] See paragraph 20 of FCC (2000), *Twelfth Report and Order, Memorandum Report and Order, and Further Notice of Proposed Rule Making*, CC Docket No. 96–45, June.

[8] The policy appears to have had an earlier impact in Oklahoma, where $1 a month Lifeline service added 6,000 new subscribers in October 2000. See Kade L. Twist, "The Digital Divide in Oklahoma Indian Country," Benton Foundation (kade@benton.org).

Povich 1997) and demand studies (Crandall and Waverman 2000; Taylor 1994). People living in wealthier and more educated communities are much more likely to have a phone in the house. Asian populations are more likely, and black and Hispanic populations less likely than white households to have a phone. Elderly populations are marginally more likely to have telephones, as are households headed by women.

Column (II) adds variables designed to capture aspects of network externalities at the local level, i.e. the idea that the household demand for telephone service depends on who else has telephone service locally. As discussed in Section 3, network externalities are a potentially important theoretical rationale for universal service policies. Controlling for population density, telephone service increases with the size of the wire center population to which the CBG belongs, suggesting that demand shifts out with the reach of local service. This stylized fact supports the hypothesis of local network externalities associated with the number of people that can be reached by a local telephone call[9]. Adding an additional 10,000 people to the wire center increases penetration by about 1/5 percentage point[10]. Controlling for population size, CBG population density reduces penetration in this regression, suggesting that face-to-face communication is to some extent a substitute for telephone usage. In contrast, Crandall and Waverman (2000), discussed in Section 4[11], find a small positive significant coefficient on population density, which they interpret as confirming a positive local network externality. Their demand analysis controlled for prices but did not include a variable for population size. Since population size and density are positively correlated, it is possible that their density variable is picking up two contrary effects, the local network externality effect, and a face-to-face communication effect (Taylor 1994 p. 236). Finally, it is noteworthy that including these variables increases the coefficient on Native American population share by several percentage points, suggesting that Native Americans tend to live in relatively unpopulated areas where the ability to make free local calls is not very valuable. Alternatively, the less negative coefficient could be an artifact of a restricted sample, which arises from the fact that wire center data are available only for large local exchange carriers.

[9] The estimated local network externality could be biased downward, because statet tariffs typically set lower prices where the number of lines is fewer. See National Association of Regulatory Commissioners (NARUC), *Bell Operating Companies Exchange Service Telephone Rates,* various years.

[10] Moreover, doubling population size and density has a significant positive effect on penetration, which is generally consistent with Perl's 1983 study, discussed in Section 3.3. Perl allowed for a nonlinear effect of phone density, and found a significant positive effect for areas with between 1,000–2,500 phones per square mile, and anegative effect elsewhere (Taylor 1994).

[11] The other studies discussed in Section 4 also include density variables ("urban" and "rural") with consistent signs.

Table 1
Determinants of CBG telephone penetration[12]

	(I)	(II)	(III)	(IV)	(V)
% Poor	−0.267	−0.259	−0.258	−0.248	−0.246
	(179.4)	(129.4)	(117.5)	(114.2)	(113.0)
Median income	0.035	0.032	0.034	0.033	0.028
	(31.1)	(22.9)	(21.6)	(21.3)	(17.6)
% Female h.o.h.	0.023	0.025	.007	−0.028	−0.033
	(12.4)	(9.8)	(2.5)	(9.9)	(11.7)
% Senior	0.004	0.003	0.006	0.002	−0.002
	(2.8)	(1.7)	(2.8)	(1.2)	(1.2)
% Children	−0.017	−0.009	−0.018	−0.012	−0.011
	(10.6)	(4.7)	(7.7)	(5.2)	(5.2)
% High school	0.117	0.103	0.111	0.102	0.098
	(63.8)	(42.7)	(41.1)	(38.5)	(36.7)
% College	0.111	0.104	0.105	0.083	0.088
	(77.8)	(57.3)	(52.0)	(40.9)	(43.1)
% Black	−0.013	−0.021	−0.009	−0.009	−0.008
	(19.8)	(22.5)	(8.8)	(9.1)	(8.1)
% Hispanic	−0.010	−0.016	−0.018	0.026	0.022
	(12.7)	(15.3)	(12.4)	(18.3)	(14.9)
% Native	−0.333	−0.247	−0.230	−0.212	−0.212
	(119.1)	(55.2)	(43.8)	(40.9)	(39.0)
% Asian	0.075	0.056	0.077	0.065	0.055
	(47.0)	(28.9)	(25.5)	(21.8)	(18.1)
% Other nonwhite	0.115	0.063	0.021	0.002	0.0010
	(5.5)	(2.11)	(0.6)	(0.5)	(2.6)
Pop. density		−0.026	−0.041	−0.40	−0.037
		(11.8)	(17.7)	(17.3)	(15.4)
W.c. population		0.020	0.019	0.012	0.013
		(37.9)	(29.6)	(18.2)	(20.2)
Loop length				−0.020	−0.016
				(19.7)	(15.5)
Average f.l. cost				−0.036	−0.033
				(41.8)	(38.0)
Controls for climate	No	No	Yes	Yes	Yes
State effects	No	No	No	No	Yes
R^2	0.537	0.531	0.551	0.564	0.580
ΔR^2					0.015
$S^2 = \frac{Var(\text{estimated state effects})}{Var(\text{telephone penetration})}$					0.017
# Observations	222,264	116,715	95,171	95,171	95,171

[12]The coefficients in this table represent the percentage point change in telephone penetration in response to a unit change in the independent variable. For example, in column (I) a 1 percentage point increase in % Poor is associated with a .267 percentage point decrease in penetration, while a $1,000 increase in Median income (which is defined in thousands of dollars) is associated with a 0.035 percentage point increase in penetration.

Column (III) controls for climate (precipitation and temperature) to capture the possibility that people living in inhospitable climates may spend more time indoors and therefore may have a greater demand for telephone service as a means of communication. This is superficially plausible, as the map in Figure 2 shows that penetration rates tend to be higher in the colder northern states. Indeed, Crandall and Waverman (2000) find a significant positive coefficient on a "cold northern state" dummy in their demand analysis. It turns out that penetration is higher where weather is more extreme[13].

Column (IV) adds FCC estimates of the monthly forward-looking cost of local service and average loop length into the mix. The argument for including these variables is that local service prices, and especially installation charges, are partly cost-based[14]. As predicted by a cost-based pricing hypothesis, higher average costs and longer loop lengths have negative effects on penetration. However, these effects are small quantitatively, as would be expected from the low price elasticities estimated by demand studies (Crandall and Waverman 2000; Taylor 1994). An extra $1 cost per month (about 3% of the mean CBG monthly cost) reduces penetration by three or four one hundreths of one percent. This implies an elasticity of about −0.01, which is roughly consistent with the demand studies under a cost-based pricing hypothesis[15]. The introduction of these supply side variables

[13] The estimated quadratic specifications for climate effects in this and subsequent regressions are:

	(III)	(IV)	(V)
Temperature	−0.5	−0.5	−0.3
	(20.2)	(21.1)	(8.0)
Temperature^2	0.004	0.004	0.003
	(17.9)	(18.6)	(8.6)
Precipitation	−0.006	−0.02	−0.03
	(−0.678)	(2.7)	(2.3)
Precipitation^2	2.56E-04	3.48E-04	4.16E-04
	(5.6)	(7.7)	(6.4)
Temp.*Precip.	−6.77E-04	−4.74E-04	−3.15E-04
	(5.1)	(3.6)	(1.6)

[14] As mentioned before, state tariffs typically set lower residential service prices in wirecenters with fewer lines, suggesting that prices are inversely related to costs within individual states. The regression, however, already captures this by controlling for the number of households served by a wirecenter. The cost-variables possibly could be picking up cost-related price variation across the states. For data on across- and within-state variation in prices see the *Bell Operating Companies Exchange Service Telephone Rates*, published annually by NARUC until 1997.

[15] Admittedly, cost-based pricing of local service is a tenuous hypothesis. Rosston and Wimmer (2000b) estimate that a 10% increase in average costs is associated with only a 0.65% percent increase in average local revenues. Such a small degree of pass-through would imply a much higher price elasticity.

does not influence the other estimated coefficients in the regression model remarkably.

Finally, column (V) includes dummy variables for the state in which the CBG is located (''state effects''). The regression indicates significant differences between states even after controlling for demography and costs. An F-test of the joint significance of the state effects easily passes, indicating that these unexplained differences between states cannot be ignored. However, the state effects adds only 0.0153 to the R^2, and the variance share of the estimated state effects (S^2) is only 0.0174[16]. Thus the state effects appear to explain somewhere between 1 and 2 percent of the variance in CBG penetration rates. These differences could be due to other population characteristics that are correlated with state of residence, or could be due to differences in state policies. Inasmuch as the total variation of penetration rates explained by the regression is not much more than 50%, the former explanation seems reasonable. However, it is unclear *a priori* what appropriate demographic or locational variables might soak up the state effects. For example, including more detailed income data into the regressions reduces the explanatory contribution of the state effects only slightly. Although it is worth entertaining the possibility that differences in state regulatory policies matter, the most optimistic interpretation of the evidence is that differences in state policies can explain no more than a small fraction of the variance in penetration rates[17].

The final regression reported in Table 1 can be interpreted as a reduced form of a structural model in which both penetration and prices are endogenous. The first equation of the structural model is a community demand curve explaining CBG penetration as a function of prices, population demographics (including proxies for network externalities), and climate, as in demand studies (Crandall and Waverman 2000; Taylor 1994). The other equations explain relevant prices as a function of access costs (proxied by loop length and forward-looking cost) and state dummies. The state dummies capture differences in state policies, e.g. different approaches to price regulation or universal service subsidies[18]. It is an open question whether price variation alone is sufficient to explain the state effects on penetration rates. Published research generally finds the price elasticity of demand for local service to be very low - on the order of -0.01 or -0.02 (Crandall and Waverman forthcoming; Taylor 1994). The price elasticity for low-income households is significantly higher (Cain and MacDonald 1991), and the elasticity

[16] S^2 is equal to the variance of the estimated state effects divided by the variance of telephone penetration. ΔR^2 is the increase in R^2 that results from adding the state effects. These two numbers can be interpreted as upper and lower bounds on the percentage of penetration variance explained by state effects. ΔR^2 is a lower bound because it implicitly attributes the explanatory power of the correlated components of the state effects to other variables. S^2 implicitly attributes the correlated components to the state effects.

[17] Sappington (2001) discusses the possibility that certain forms of incentive regulation may increase penetration rates.

[18] Differences in state universal service policies, which establish low-income subsidies, are discussed later.

with respect to installation charges is significantly higher than for monthly service charges (Hausman, Tardiff, and Belinfante 1993; Crandall and Waverman 2000). Thus published economics research finds some weak support for universal service policies that target low income households and focus on lowering installation charges. These are the aims of the FCC's Lifeline and Link-Up programs, which are evaluated in Section 4.

An intriguing possibility is that some of the substantial unexplained geographic variation in penetration rates is due to "coordination failures" associated with network externalities[19]. The basic economics of the telephone network externality is that an individual subscriber benefits when other consumers connect to the network. This interdependence of decision-making creates a coordination problem for consumers: "If enough consumers connect, then so will I, but if others don't connect then neither will I." Thus, under the network externality hypothesis, consumer decision-making depends on consumers' expectations about other consumers' decision-making. The circular reasoning inherent in consumer coordination problems allows multiple equilibria, e.g. low level equilibria in which few people connect to the network, and high level equilibria in which many connect. Depending on nonlinearities in demand, there can be many equilibria for a given community, yielding a variety of different possible stable penetration levels. Thus, in theory, part of the geographic variation in penetration levels could be due to similar communities arriving at different equilibrium levels of penetration for historical reasons. The significance of network externalities for optimal telecommunications pricing is discussed further in Section 3 below.

The questions "Could the United States do more to promote universal service?" and "Do state policies matter for the achievement of universal service goals?" are important questions in the realm of positive economics. The corresponding normative questions are "What are optimal levels of telephone penetration and how do they vary with the characteristics of consumer groups?" and "What are the best ways to achieve universal service goals?" The next section surveys what economic theory has to say about these and related normative questions.

3. Normative economics of universal service

3.1. Price distortions

Perhaps the most fundamental advice of economists is that marginal cost pricing maximizes economic efficiency. As discussed in detail in following subsections, the standard marginal cost pricing prescription must be qualified in the presence of

[19] See Katz and Shapiro (1994) and Liebowitz and Margolis (2001) for discussions of network effects.

scale economies and network externalities. Nevertheless, economists generally agree that universal service policies that distort usage prices above incremental costs sacrifice economic efficiency.

In the United States, access regulation and universal service policies have helped keep the prices of long distance usage above marginal cost. For example, the price of an interLATA long distance call carried by AT&T reflects federally-mandated access charges paid to the local telephone companies who originate and terminate the call. Almost everyone recognizes that usage-based components of these access charges have been maintained above the marginal cost of access[20]. Hausman (1998) and Prieger (1998) interpret the resulting price distortion as a usage tax[21], and use approximations from public finance theory to measure the resulting loss of economic efficiency[22]. The analysis below follows Hausman's logic closely, but measures efficiency losses exactly by assuming a constant elasticity of demand over the relevant range.

The basic issue is illustrated in Figure 3, adapted from Hausman (1998). The price per minute of long distance is p, the marginal cost is c, and usage is q. The usage tax is t. In the absence of the tax, consumers would pay $p - t$ per unit of long distance usage. The revenue raised from the tax is

$$R = tq \tag{1}$$

For an otherwise fixed market structure, the efficiency loss from the tax (called "deadweight loss" by economists) is measured by the sum of areas A and B. Area A represents the reduction in profits ("producer surplus") caused by the tax, assuming the tax is fully passed on to consumers[23]. Area B is the loss of consumer welfare ("consumer surplus") from the tax.

The deadweight loss per unit of tax revenue raised can be calculated as follows. Assume that the demand for long distance usage has a constant elasticity ε over the range of prices between $p - t$ and p. Then the reduction in quantity resulting from the usage tax is

$$\Delta q = \left[\left(1 - \frac{t}{p}\right)^{-\varepsilon} - 1\right] q \tag{2}$$

[20] The FCC is phasing out significantly above-cost usage-based access prices, replacing them with higher fixed charges and with revenue-based universal service "contributions" (i.e. revenue taxes).

[21] The FCC is moving from a system of usage taxes, implicit in access taxes, to a system of revenue taxes, implicit in the calculation of universal service contributions. Depending on market structure, revenue taxes may be more efficient than usage taxes.

[22] Hausman (2001) applies the methodology to the market for mobile telephony. See additional references therein.

[23] The assumption of full pass through is hard to defend theoretically in an oligopoly context, and exaggerates the efficiency loss if the tax partially extracts rents from oligopoly market power. Further analsysis of tax incidence and welfare consequences in the oligopoly case would clarify the debate on efficiency losses from usage price distortions.

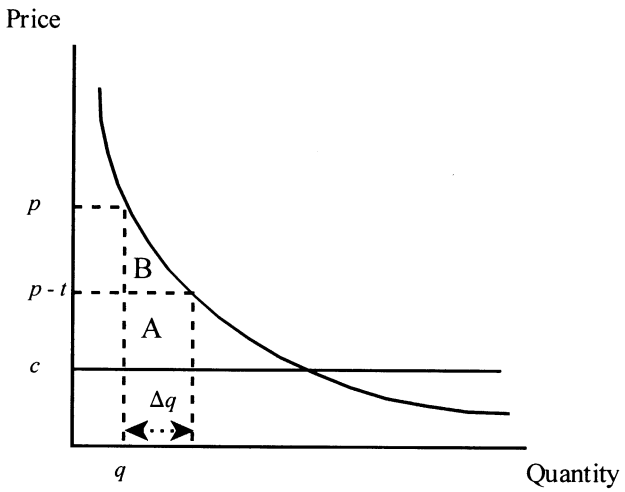


Fig. 3. Consequences of an access tax.

and loss of producer surplus (Area A) is

$$(p - t - c)\Delta q = (p - t - c)\left[\left(1 - \frac{t}{p}\right)^{-\varepsilon} - 1\right]q \tag{3}$$

The corresponding loss of consumer surplus (Area B) is calculated by integrating the demand curve between $p - t$ and p and subtracting tax revenue. This gives the formula

$$\left\{\frac{1}{1-\varepsilon}\left[p - (p - t)\left(1 - \frac{t}{p}\right)^{-\varepsilon}\right] - t\right\}q \tag{4}$$

The incremental loss of economic efficiency (''incremental deadweight loss'') is equal to the sum of lost producer surplus and consumer surplus. Simple calculations yield an expression for the incremental deadweight loss per unit of tax revenue: adding the expressions for lost producer and consumer surplus in Equations (2) and (3), and dividing by the definition of tax revenue in Equation (1), yields lost consumer and producer surplus per unit revenue raised by the tax; adding these up reveals that the average incremental deadweight loss equals

$$\left(\frac{p}{t} - 1 - \frac{c}{p}\frac{p}{t}\right)\left[\left(1 - \frac{t}{p}\right)^{-\varepsilon} - 1\right] + \left\{\frac{1}{1-\varepsilon}\left[\frac{p}{t} - \left(\frac{p}{t} - 1\right)\left(1 - \frac{t}{p}\right)^{-\varepsilon}\right] - 1\right\}.$$

The significance of this complicated-looking formula is that a calculation of the average incremental deadweight loss from the price distortion caused by the

access tax requires three numbers: the demand elasticity ε, the tax rate $\frac{t}{p}$, and the cost share $\frac{c}{p}$. Some representative calculations are presented in Table 2. Each entry in the table has two numbers. The first (larger) number is the incremental deadweight loss per unit of tax revenue; the second number is the corresponding loss of consumer surplus alone. A consensus estimate of the price elasticity of long distance usage is about $\varepsilon = 0.7$ (Taylor 1994). For this elasticity, if the tax rate and cost shares are $\frac{t}{p} = 0.25$ and $\frac{c}{p} = 0.25$, the incremental deadweight loss is \$0.55 per unit of revenue[24], of which \$0.10 is lost consumer surplus, the rest being lost profit. In other words, every dollar of revenue raised by the tax costs the economy an additional fifty-five cents and reduces consumer surplus by ten cents[25].

A debatable aspect of this analysis is the calculation of lost producer surplus. Hausman's calculations make sense if there are prohibitive barriers to entry into the long distance market, enabling incumbent firms to sustain supracompetitive profits. In this case, elimination of the tax does not cause a change in market structure, and area A represents an increase in industry profits that results from the expansion of incumbent firms. However, as Hausman (1998) notes, it is possible "that the industry is imperfectly competitive and price exceeds marginal cost to cover fixed costs." In this case, the elimination of the tax could prompt additional entry, and at least part of area A represent the additional fixed costs incurred by the new entrants. Increased industry fixed costs do not add to economic welfare, suggesting that Hausman's calculation of the efficiency loss from an access tax is biased upward. Indeed, if equally efficient firms drive equilibrium profits to zero both before and after the elimination of the tax, then the efficiency loss from the access tax is only the loss in consumer surplus measured by area B, which is the second, smaller number in each entry of Table 2[26]. Thus one's

[24] Perhaps surprisingly, the average efficiency loss is not monotonic in $\frac{t}{p}$. This is because an increase in $\frac{t}{p}$ increases both numerator (total efficiency) and the denominator (tax revenue) of the expression for average efficiency loss.

[25] Hausman (1998) apparently estimated the deadweight loss using a second-order Taylor series approximation, although his precise calculations are difficult to unravel. He also assumed a higher tax rate of $\frac{t}{p} = .403$, which was plausible a few years ago before price caps lowered access rates. He arrived at an estimated deadweight loss of \$0.654 for each dollar of revenue raised. Substituting $\frac{t}{p} = .403$ into the above exact formula yields a smaller \$0.56. Prieger (1998) applies a similar public finance methodology (and explains it better) to estimate the deadweight loss from prospective universal service taxes. The point is the same. Price distortions to support universal service potentially entail substantial efficiency losses. The authors agree that a more efficient way to fund explicit universal service subsidies would be to tax local access. See also Hausman (1999).

[26] More generally, if entry is "lumpy", then abnormal long run profits can persist in a free entry equilibrium. However, it is unclear *a priori* whether industry profits will rise or fall if the elimination of a tax prompts additional entry. If industry profits were to fall then the efficiency loss from the tax would be even less than area B, and conversely. Lacking finely detailed information on market structure, it appears reasonable to assume a zero effect of entry on long run industry profits and to measure the efficiency loss by area B alone. However, if firms differ in efficiency, then part of area A could represent the rents of the more efficient firms, in which case the efficiency loss per unit of tax revenue is somewhere between the two numbers reported in Table 2.

Table 2
Efficiency and consumer surplus losses per $ tax revenue

	$\frac{c}{p}=0$		$=0.25$		$=0.50$		$=0.75$	
$\varepsilon=0.6$								
$\frac{t}{p}=0.25$	0.65	0.09	0.46	0.09	0.28	0.09	0.09	0.09
$=0.50$	0.73	0.21	0.47	0.21	0.21	0.21		
$=0.75$	0.85	0.42	0.42	0.42				
$\varepsilon=0.7$								
$\frac{t}{p}=0.25$	0.77	0.10	0.55	0.10	0.33	0.10	0.10	0.10
$=0.50$	0.88	0.25	0.56	0.25	0.25	0.25		
$=0.75$	1.06	0.51	0.51	0.51				
$\varepsilon=0.8$								
$\frac{t}{p}=0.25$	0.90	0.12	0.64	0.12	0.38	0.12	0.12	0.12
$=0.50$	1.04	0.29	0.67	0.29	0.29	0.29		
$=0.75$	1.29	0.61	0.61	0.61				

perspective on the efficiency loss from universal service taxes depends on assumptions about the industrial organization of the long distance market[27].

Hausman (1998 p. 14) argues that a more relevant calculation is the marginal effect of reducing usage taxes. Hausman assumed that any increase in the usage tax is fully passed on to consumers. Under this assumption, the marginal deadweight loss with respect to t is

$$\left(1-\frac{c}{p}\right)\varepsilon q \tag{4}$$

of which

$$\varepsilon\frac{t}{p}q \tag{5}$$

is the marginal loss in consumer surplus. The marginal tax revenue for an increase in t is

$$\left(1-\varepsilon\frac{t}{p}\right)q. \tag{6}$$

[27] Prieger (1998 p. 66) recognizes that the efficiency loss depends on industry structure, but downplays it by suggesting that short run entry barriers might allow above-normal profits to persist temporarily. His calculations (1998 Table 2) confirm that the welfare loss from an access tax is much lower in the long run once new entry erodes the temporary market power of the incumbents. See Kaserman and Mayo (2001) for a detailed discussion of the industrial organization of the long distance market.

Dividing (4) and (5) by (6) gives the marginal efficiency loss and the marginal consumer surplus loss for an extra dollar of tax revenue raised by an increase in the usage tax. Table 3 presents some representative calculations. Following Hausman, these calculations assume that an increase in the usage tax is fully passed on to consumers in the final price. For example, if $\varepsilon = 0.7$, $\frac{t}{p} = 0.25$ and $\frac{c}{p} = 0.25$, then a \$1 increase in the amount of revenue raised by the access tax costs society an additional \$0.64, of which \$0.21 is a direct loss to consumers. A comparison of Tables 2 and 3 shows that marginal losses exceed average losses.

Hausman argues that it would be more efficient to finance universal service subsidies from general tax revenues. He bases this recommendation on published estimates of the marginal efficiency losses of general taxes ranging between 0.260 and 0.395 (Hausman 1998 p. 15). Table 3 shows that the marginal welfare effects of the asset tax exceed this range (for $\varepsilon = 0.70$) if lost producer surplus (area A of Figure 3) is part of incremental deadweight loss. However, if producer surplus is dissipated by entry costs, as in a symmetric free entry oligopoly equilibrium, then the marginal welfare effect of the usage tax, which is equal to the marginal consumer surplus loss, is less and may be below the marginal social cost of public funds. Thus, depending on the industrial organization of the long distance market, the access tax may or may not be an economically attractive method to finance universal service compared to financing out of general revenues[28].

Table 3
Marginal efficiency and consumer surplus losses

	$\frac{c}{p} = 0$		$= 0.25$		$= 0.50$		$= 0.75$	
$\varepsilon = 0.6$								
$\frac{t}{p} = 0.25$	0.71	0.18	0.53	0.18	0.35	0.18	0.18	0.18
$= 0.50$	0.86	0.43	0.64	0.43	0.43	0.43		
$= 0.75$	1.10	0.82	0.82	0.82				
$\varepsilon = 0.7$								
$\frac{t}{p} = 0.25$	0.85	0.21	0.64	0.21	0.42	0.21	0.21	0.21
$= 0.50$	1.08	0.54	0.81	0.54	0.54	0.54		
$= 0.75$	1.47	1.11	1.11	1.11				
$\varepsilon = 0.8$								
$\frac{t}{p} = 0.25$	1.00	0.25	0.75	0.25	0.50	0.25	0.25	0.25
$= 0.50$	1.34	0.67	1.00	0.67	0.67	0.67		
$= 0.75$	2.00	1.50	1.50	1.50				

[28] The industrial organization literature recognizes that oligopoly entry may be excessive from a social perspective (Mankiw and Whinston 1986). In this case, an access tax can improve social efficiency by reducing excessive entry.

Hausman's main policy recommendation is that universal service is best achieved by targeted subsidies financed by a fixed universal service tax on access. The FCC is moving in this direction by reducing per minute long distance access charges and by raising the monthly subscriber line charge (SLC). The wisdom of "going all the way" and completely eliminating per minute access charges depends on scale economies and network externalities, discussed in the next two subsections.

3.2. Scale economies

Local economies of scale provide a rationale for universal service policies, although this economic argument does not feature prominently in today's policy debates on the subject. Certainly, local scale economies cannot be dismissed out of hand. Maher (1999) reports modest estimated scale economies in access, based on central office cost data provided anonymously by two local telephone companies. If there are economies of scale of connecting people, then adding people to the network lowers the average cost of connections, potentially to the benefit of all.

The *a priori* plausibility of local scale economies depends on the nature of the universal service problem. One flavor of scale economy is an economy of density. An increase in telephone penetration at a wire center service area that is already built out amounts to an increase in the number of lines served in a given geographic area. For example, if 95 out of 100 households on a street already are getting telephone service, then the incremental cost of serving an additional household must be less than the average incremental cost of serving the street. The reason is that the necessary poles and conduits, and perhaps even spare copper wire pairs, are already in place. Thus scale economies are very plausible if the universal service problem is to increase penetration in a given service area.

Another flavor of local scale economy is an economy of geographic scope. If greater penetration requires extending the perimeter of the wire center, then it is plausible that the incremental cost of service is either greater or less than the average cost. On the one hand, average cost may decline because the geographic extension relies on existing remote terminals, transport and switching infrastructure. On the other hand, the greater costs of installing and maintaining longer copper wire loops could cause the incremental costs of service to rise above the average cost. For this reason, economies of geographic scope seem less plausible than economies of density as a source of local scale economies.

The economies of scale rationale for universal service poses a well known dilemma. Average cost pricing results in an inefficiently low level of penetration, but marginal cost pricing leaves a deficit to be funded somehow. What's a regulator to do? The famous Ramsey rule for second-best pricing resolves the dilemma optimally by marking-up prices above marginal cost in inverse proportion to the price elasticity of demand.

Most U.S. households pay a fixed monthly price for access (and local service) and usage sensitive prices for long distance calling. The long distance prices may

depend on whether the call is intrastate or interstate, and on the distance of the call. However, a simple two-part service arrangement featuring a fixed usage price provides a good basis for an analysis of optimal pricing with economies of scale. The standard Ramsey rule requires some modification if there are separate prices for access and usage. The modification is required because access is a necessary ingredient of residential access to the telephone network. This section outlines the relevant theory of optimal two-part tariffs, along the lines developed by Brown and Sibley (1986), Vogelsang and Mitchell (1991), and Schmalensee (1981). It is appropriate to interpret the economy of scale in the theoretical model as an economy of density.

Ramsey pricing rules are based on demand as well as costs. Thus, the derivation of the optimal pricing rule requires a model of both. To keep matters simple, assume that there are just two services, usage and access, and a separate price for each, p and r[29]. Consumer heterogeneity is represented by a parameter θ. A type θ individual has a utility (consumer surplus) of

$$V = U(p, \theta) - r$$

from connecting to the telephone network. A service plan that is more favorable to the consumer yields a higher consumer surplus.

Different types of consumers have different preferences over service plans. To simplify further, assume a multiplicatively-separable functional form

$$U(p, \theta) = \theta u(p),$$

Thus the consumer surplus of a type θ consumer with service plan (p, r) is

$$V = \theta u(p) - r$$

where $u(p)$ is assumed to be a smooth, convex, and decreasing function of p. A consumer with a higher value of θ is more willing to accept a higher access price for a lower usage price. However, all consumers have the same price elasticity of demand for usage. By a standard economic argument[30], a type θ individual has a demand curve for usage,

$$\begin{aligned} X(p, \theta) &= -\theta u'(p) \\ &\equiv \theta x(p), \end{aligned}$$

[29] This is an oversimplification: usage can be interpreted as long distance usage, with local usage bundled into access. More generally, economic efficiency requires separate usage-sensitive prices for local and long distance usage, because these have different price elasticities.

[30] The argument is known in the consumer theory literature as Roy's identity. The partial equilibrium framework adopted here assumes a constant marginal utility of income, implicitly interpreting a decrease in r as an increase in income.

that is derived from the utility function. The corresponding price elasticity of demand for usage is

$$\epsilon = -\frac{px'(p)}{x(p)}$$

The price elasticity might depend on p, but it does not depend on θ.

Only consumers with a positive consumer surplus will opt to connect to the network. The marginal type is θ_o satisfying

$$r = \theta_o u(p), \tag{7}$$

meaning that this consumer is just indifferent between connecting or not. By substituting this expression for r into the utility function expressed by Equation (7), the consumer surplus of a type θ is written as

$$V = (\theta - \theta_o)u(p),$$

which is a function of the consumer's type, the marginal consumer type, and the usage price.

If M is the total number of consumers, and N are connected, then the penetration rate is $n = \frac{N}{M}$. The penetration rate is related to the identity of the marginal consumer by the formula

$$n = \int_{\theta_o}^{\infty} f(\theta)d\theta \equiv [1 - F(\theta_o)]$$

Here $f(\theta)$ is the frequency (density) of type θ consumers in the population, and $F(\theta)$ is the fraction of consumers who make fewer calls than does a type θ. In this model, the elasticity of the penetration rate with respect to the access price r is

$$\eta = \frac{\theta_o f(\theta_o)}{n}$$

This "access elasticity" measures the sensitivity of the marginal consumer to a change in the access price. The average consumer is type

$$\bar{\theta} = \frac{\int_{\theta_0}^{\infty} \theta f(\theta)d\theta}{n},$$

makes $\bar{\theta}x(p)$ calls, and enjoys a consumer surplus of $\bar{\theta}u(p) - r$. The average consumer surplus over the entire population is therefore

$$\bar{V} = n[\bar{\theta}u(p) - r]. \tag{8}$$

Substituting Equation (7) for the marginal consumer into (8) yields an expression for average population consumer surplus,

$$\bar{V} = n(\bar{\theta} - \theta_o)u(p)$$

as a function of the usage price, the marginal consumer, and the average consumer.

Now turn to costs and profits. Assume for simplicity that the marginal cost of usage is a constant at c. The average cost of a connection is $\bar{h}(\theta_o)$ when all types $\theta \geqslant \theta_o$ are connected to the network. The marginal cost is related to the average cost according to the formula

$$h(\theta_o) = \bar{h}(\theta_o) - \frac{\bar{h}'(\theta_o)\theta_o}{\eta}.$$

An economy of scale in providing access exists if $\bar{h}'(\theta_o) > 0$. In this case the marginal cost of a connection is lower than the average cost. This means that, as more subscribers are added to the network, the average cost declines. The profits earned on an average consumer are $r + (p - c)\bar{\theta}x(p) - \bar{h}(\theta_o)$. Using Equation (7) to substitute for r and averaging over the entire population yields an expression for the average population profit,

$$\bar{\Pi} = n[\theta_o u(p) + (p - c)\bar{\theta}x(p) - \bar{h}(\theta_o)].$$

The problem of maximizing total welfare subject to a break-even constraint on profits amounts to maximizing a weighted sum of consumer surplus and profits according to the Lagrangian function

$$L = \bar{V} + (1 + \lambda)\bar{\Pi}$$

where $\lambda \geqslant 0$ is the shadow price of the break-even constraint. In other words, the optimal service plan maximizes an appropriately weighted sum of consumer surplus and profits (producer surplus). The greater weight on profits reflects the cost to society of solving the Ramsey dilemma of how best to recover access costs. As shown below, the shadow price is strictly positive if there are economies of scale.

Maximizing L with respect to p yields the modified Ramsey formula for pricing usage (Brown and Sibley 1986 p. 95):

$$\frac{p - c}{p} = \frac{\lambda}{1 + \lambda}[1 - \varpi]\frac{1}{\epsilon}$$

where ϵ is the price elasticity of usage defined earlier, and the variable

$$\varpi = \frac{\theta_0}{\bar{\theta}}$$

is equal to the ratio of usage of the marginal consumer to average usage. Thus, assuming that the Ramsey dilemma is real and $\lambda > 0$, the usage markup is higher the greater is the difference in usage between the marginal and average subscriber. Brown and Sibley (1986 p. 96) interpret $1 - \varpi$ as "an adjustment term accounting for the cross-elasticity between consumption and participation." More specifically, the adjustment accounts for the facts that an increase in p requires a decrease in r in order to maintain penetration, and that this rebalancing impacts both average utility and profits.

The usage formula makes clear that marginal cost pricing can solve the welfare maximization problem only if $\lambda = 0$. This case obtains for a particular value of θ_0. In this singular case $p = c$, requiring $r = \bar{h}(\theta_o)$ if the firm is to break even. This consumer type is just willing to accept a strictly cost-based service plan with access price $r = \bar{h}(\theta_o)$ and a usage price $p = c$. Can this be optimal? The answer is no if there is an economy of scale in connecting people to the network, i.e. if

$$\bar{h}'(\theta_o) > 0$$

To reach this conclusion, consider how social welfare changes with the identity of the marginal consumer. Evaluating the derivative of L with respect to θ_0 at the point of strict cost-based pricing yields

$$\frac{dL}{d\theta_o} = -\bar{h}'(\theta_o)n$$

which is unambiguously negative if $\bar{h}'(\theta_o) > 0$, meaning that welfare would be increased by lowering θ_o. But then the profit constraint becomes binding, i.e. $\lambda > 0$, and $p > c$ according to the usage formula. Thus, economies of scale provide a clear rationale for "price distortions." The average consumer benefits from the resulting network expansion because economies of scale enable a lowering of the access price relative to the increase in the usage price.

The optimal access price satisfies a modified Ramsey formula that appropriately accounts for opportunity costs. The first-order condition for optimal θ_o yields (Brown and Sibley 1986 p. 95)

$$\frac{r - m}{r} = \frac{\lambda}{1 + \lambda}\frac{1}{\eta}$$

where

$$m = h(\theta_o) - (p - c)\theta_o x(p)$$

is the marginal opportunity cost of a connection. The formula modifies the standard Ramsey inverse elasticity rule by treating marginal usage revenues as a component of marginal opportunity cost. A key observation from the formula is that, for purposes of optimal access pricing, the theoretically correct definition of marginal cost is marginal opportunity cost, which subtracts the usage profits earned on the marginal consumer from the marginal cost of a connection.

Economists' advice that usage should be priced close to its marginal cost is based on empirical evidence that the access elasticity is small, and on an implicit assumption that the revenue contribution of the marginal consumer is not likely to be large relative to marginal cost[31]. For example, suppose that the usage profits on the marginal consumer just cover the marginal cost of a connection. Then $m = 0$, $\frac{\lambda}{1+\lambda} = \eta$, and $\frac{p-c}{p} = (1 - \varpi)\frac{\eta}{\varepsilon}$. If the access elasticity (η) is small relative to the usage elasticity, then the usage markup is small. Empirical estimates of the price elasticities of access and (long distance) usage are in the neighborhood of $\eta = 0.02$ and $\varepsilon = 0.70$, i.e. the usage elasticity is an order of magnitude greater than the access elasticity, which implies that the usage markup is small. Thus, unless the profit contribution of marginal consumers exceeds the marginal connection cost significantly, scale economies do not appear to be an important justification for large price distortions to achieve universal service[32].

3.3. Network externalities

Network externalities are inherent in the idea of a telephone network. The larger the network, the more people there are to call, and therefore the greater is the value of being connected to the network. Although network externalities provide a clear rationale for universal service policies, it is a rationale that has lost center

[31] The fact that penetration is lower for lower income households suggests that marginal consumers are predominantly lower income households. Crandall and Waverman (2000) document that lower income households do spend less on long distance usage, although the difference is not a dramatic one.

[32] This conclusion needs some qualification. If the average demand is great, then even a small usage markup (i.e. small λ) can justify a significant access discount. Moreover, it is possible to construct realistic examples of optimal two-part tariffs featuring both small usage markups and moderate access discounts. Using a model calibrated to 1970 data Mitchell (1978 p. 531) calculated that the optimal two-part tariff for local service has moderately-sized access discounts and usage markups, while achieving a high penetration rate. However, it is noteworthy that price elasticity for local usage implicit in Mitchell's model is significantly less than the consensus 0.7 elasticity for long distance usage (Mitchell 1978 p. 528). Building on Mitchell's example, Brown and Sibley (1986 p. 96) calculated that the optimal two-part tariff raised average consumer welfare by 5 cents a month compared to pricing usage at average cost, although at the cost of significantly reduced penetration.

stage in the policy debate. Laffont and Tirole (2000 p. 230) offer the following explanation for its neglect:

> Network or club externalities are no longer at the forefront of the universal service debate (except perhaps for new services such as the Internet), partly because networks are largely developed in OECD countries and partly because it is recognized that network externalities are to a large extent internalized by operators.

This dismissal of the network externality rationale for universal service is not fully convincing. The argument that network externalities are unimportant in developed economies rests on an assumption that the average subscriber to the network does not have much interest in calling the marginal subscriber[33]. Crandall and Waverman (2000 p. 25) put the argument this way:

> (T)he network externality argument has little relevance for telephony in developed economies today for several reasons. If my telephone in Manhattan reaches 2 million people, another connection will probably have little value to me. Of course, if that connection is my mother, then the connection is of real value to me, and ... I can subsidize her telephone directly! Otherwise, there is no reason why I – in Manhattan – should subsidize someone in Kalamazoo.

The rhetoric does not quite hit its mark. Even if the average telephone subscriber in Manhattan places a small value on being able to call the marginal subscriber, multiplication of that small value by 2,000,000 can be a large number. Moreover, there surely are people in Manhattan who value calling people in Kalamazoo; that is, a network externality can be long distance as well as local. The magnitude of the network externality remains an empirical issue on which evidence is scant.

How do regulated firms internalize network externalities? To a large extent, this is up to the regulators. Raising the price of usage above its marginal cost, and reducing the price of access below its incremental cost, encourages the subscription of consumers who most likely do not originate a lot of calls[34]. Nevertheless, these subscribers may receive calls from other consumers who benefit from making these calls. Moreover, the increased call volume from this externality generates additional revenue which limits the need to raise the usage price to cover the access deficit. The economic efficiency of such price distortions is the focus of the network externality debate.

It is not hard to construct a theoretical model that illustrates the potential importance of network externalities. Consider a telephone network serving N consumers. Suppose that each consumer is potentially interested in calling a

[33] There are other less obvious network externalities. A large subscriber base creates a "market" for various network-based transactions. e.g. bank by phone. Such indirect network externalities most likely are less important for mature networks, but arguably of crucial importance for emerging networks such as the Internet. See Katz and Shapiro (1994).

[34] On the other hand, Hausman et al. (1993) report estimates suggesting that rebalancing rates in the opposite direction could increase penetration.

fraction θ of the others, and places an average of $x(p)$ calls to each at a price of p. Therefore, the number of calls the consumer makes is

$$X(p, \theta, N) = \theta(N-1)x(p)$$

and the consumer's value of calling is

$$U(p, \theta, N) = \theta(N-1)u(p).$$

where the relationship between $u(p)$ and $x(p)$ is as in the previous section. Each consumer's usage and value of being connected increases linearly with the number of other consumers connected to the network. This is a mathematical statement of a particularly strong network externality[35]. More generally though, the network externality hypothesis only requires that value increases monotonically with subscribers.

The consequences of network externalities for usage prices can be derived by building on the previous model of optimal two-part pricing; see Vogelsang and Mitchell (1991) for a literature survey and a related model. With a network externality, the utility of an average subscriber is

$$\bar{\theta}(N-1)u(p) - r$$

and the marginal consumer (type θ_o) is defined by

$$r = \theta_o(N-1)u(p).$$

Substitution and multiplication by the penetration rate gives the population average utility,

$$\bar{V} = n(\bar{\theta} - \theta_o)(N-1)u(p).$$

Similarly, if the average network cost is fixed at $\bar{h}$ (ignoring scale economies), then the population average profit is

$$\bar{\Pi} = n\{(N-1)[\theta_o u(p) + (p-c)\bar{\theta}x(p)] - \bar{h}(\theta_o)\}$$

The Lagrangian is defined as before, and the "Ramsey formula" for the optimal usage price is exactly the same:

$$\frac{p-c}{p} = \frac{\lambda}{1+\lambda}[1-\varpi]\frac{1}{\epsilon}$$

[35] This is a statement of "Metcalfe's Law" that the value of a network increases with the number of users squared. Robert Metcalfe was the founder of 3Com Corporation.

where ϖ is the ratio of marginal to average usage. The optimal access price generalizes the previous formula:

$$\frac{[r-m]}{r} = \frac{\lambda}{1+\lambda}\frac{1}{\eta} - \chi;$$

with opportunity cost similar to as before:

$$m = \bar{h} - (p-c)(N-1)\theta_o x(p);$$

and a new term reflecting the network externality:

$$\chi = \frac{N-1}{N}\left[\frac{\bar{h}}{r} + \frac{1}{1+\lambda}\left(\frac{1}{\varpi} - 1\right)\right].$$

As in the case of scale economies, marginal cost pricing is not optimal, i.e. $\lambda > 0$, which requires an access deficit ($r < \bar{h}$) from the break-even constraint.

The network externality clearly justifies pricing access below the average cost of a connection, perhaps substantially depending on the values of λ and ϖ. Using the approximation $\frac{N-1}{N} \approx 1$, and assuming constant returns to scale, we obtain

$$\frac{r-\bar{h}}{r} \approx \frac{1}{1+\lambda}\left(\frac{1}{\varpi}\right)^2 - \frac{\lambda}{1+\lambda}\left(\frac{1}{\eta} - 1\right)\left(\frac{1}{\varpi}\right)$$

The following is an example demonstrating that the network externality can justify a significant discount on the price of access. Suppose that the elasticities of access and usage are $\eta = 0.02$ and $\varepsilon = 0.70$, respectively, and that costs are $c = 0.015$ and $\bar{h} = 20$ with no scale economies. In dollars and cents, this means that the marginal cost of usage is a penny and a half, and the cost of access is \$20. Suppose further that $\varpi = 0.11$ and $\lambda = 0.186$[36]. The solution to the model for this example is: $p = 0.019$; $r = 16.90$. The optimal usage price is just under two cents and the optimal access price is just under \$17.

The example demonstrates that the network externality hypothesis potentially provides a sound theoretical rationale for subsidizing access to achieve universal service. Of course the model is too simple for practical purposes and probably overstates the case for an access subsidy. One blemish is the unrealistic assumption that doubling the size of the network also doubles the amount of

[36] These values can be justified by a suitable choice of distribution function for θ, and by a suitable multiplicative scaling of the value of usage. The usage ratio $\varpi = 0.11$ determines the penetration rate from the distribution of types; for example, if θ has a standard uniform distribution, then the implied penetration rate is about 94%. The Lagrange multiplier $\lambda = 0.186$ means that it costs the economy an additional \$0.18 for every \$1.00 raised this way via the usage markup.

usage at a given price. Telephone calls take time and consumers have other things to do. An increasing opportunity cost of time will curtail telephone usage even as network size grows. Nevertheless, with more calling opportunities, consumers can substitute from lower to higher value calls. The increased substitution opportunities of a larger network still validates the network externality hypothesis even if consumers do not make more calls. However, the rising opportunity cost of calls does lessen the quantitative significance of the network externality.

A second blemish is that the model assumes that all consumer types receive the same number of calls, even though they differ in their originating usage. It is possible and perhaps likely that people who make few calls when connected to the network also tend to receive few calls. The external benefits of connecting such people to the network are small. If this were true for marginal users as a class, then the case for an access subsidy is weakened significantly. This apparently is what Crandall and Waverman mean in the quotation above. However, the empirical validity of this intuitively plausible hypothesis remains unclear.

A final blemish is that the analysis ignores call externalities. A call externality occurs when some of the benefits of a telephone call accrue to the recipient, and are not internalized by the caller. In the United States and elsewhere the calling party pays for the telephone call[37], and may decline to place a call if the price is too high, even though the joint benefits of the call are worth the cost. For example, I may wait for you to call me, and vice versa, and the call gets put off. The model can be modified to account for call externalities by supposing that the value of receiving a call is (on average) equal to v. Then the Ramsey formula for optimal usage pricing becomes

$$\frac{p-c}{p} = \frac{\lambda}{1+\lambda}\left[(1-\varpi)\frac{1}{\epsilon} - \frac{v}{p}\right].$$

Clearly, if v is sufficiently large relative to p, then the optimal markup is negative, i.e. it is optimal to encourage more calls by setting the usage price below marginal cost[38]. Clearly, if usage is priced below cost, then access must be priced above cost if the firm is to break even. Thus, the call externality could completely undermine the case for access subsidies based on scale economies and network externalities.

[37] An exception is a call to a wirelesss phone. In the United States the wireless receiver pays airtime charges. Elsewhere in the world, "calling party pays" is the norm even for wireless calls, and there is a move afoot for the FCC to require a "calling party pays" option in the U.S. as well.

[38] Note that $\frac{\lambda}{1+\lambda}v$ can be interpreted as an additional component of opportunity cost in the Ramsey formula for usage prices. The reason for the $\frac{\lambda}{1+\lambda}$ adjustment is that the call externality enables the firm to charge a higher access price to the marginal consumer.

Despite the conflict between call externalities and network externalities, the former has not received as much attention in the academic literature. Brown and Sibley (1986 p. 197) put the case against call externalities this way:

> The call externality is probably not too important. It only involves two people and can probably be easily "internalized." For example, two frequent callers could arrange to share the cost of calling. Furthermore, not all call externalities are positive externalities; there are certain phone calls that one is annoyed to receive. Since the telephone company cannot be expected to distinguish between positive and negative call externalities, it is probably not useful to incorporate them into price formulas. For this reason, and because call externalities can be internalized fairly well, they do not provide a strong case for call price reductions.

Vogelsang and Mitchell (1991) give more credence to the call externality by observing that successful bargaining over how to divide the cost of calling may itself require a costly telephone call. They also argue that call externalities are relatively more important in developed economies; their reason is that call externalities involve interactions among all consumers, while network externalities only involve interactions with marginal consumers. In the context of the above theoretical model, this means that, while network externalities increase with network size at rate N, call externalities increase at rate N^2. This is an interesting theoretical argument. However, empirical evidence on the relative significance of call and network externalities is lacking.

There are scraps of evidence on network externalities in telecommunications networks. As discussed in Section 2, Crandall and Waverman (2000) find a positive effect of population density on the demand for residential access, and interpret this as supporting the network externality hypothesis. Another scrap of evidence comes from Louis Perl's 1983 unpublished study of access demand, summarized by Taylor (1994 p. 86–96). Perl included in his discrete choice model measures of the size and density of the local network. His estimates imply that doubling size and density of a local network of 25,000 increases the average value of a subscription by $4.36, while doubling the network again creates another $1.17 of value for each subscriber (Taylor 1994 pp. 236–8). Thus, only modest network externalities appear at the local level, and the magnitude of the local network externality declines with size.

Network externalities can be either local or long distance. It is valuable to reach more people with a long distance call, as well as to be able to place more calls within a local service territory. It is unclear *a priori* which kind of network externality is the more important. The value of being able to call someone on the telephone depends both on the price of the call and on the availability of alternative means of communication. On the one hand, even though a local call typically is free, face-to-face communication is often an excellent alternative. On the other hand, a long distance call, while costly, often lacks a good substitute.

The fact that long distance prices have been dropping sharply suggest that long distance network externalities are becoming more important[39].

The network externality hypothesis allows that usage increases with the number of connected consumers. Taylor (1994 Appendix 3) estimated a log linear equation relating the average number of calls from city A to city B to relevant prices, the average household income in A, and the number of addressable telephones in B (market size) using quarterly data on off-peak long-haul traffic between Canadian cities between 1974 and 1983. The estimated elasticity of usage with respect to market size was 1.482 with a t-value of 8.5! It is not clear what to conclude from this estimate. Taylor speculates that the high elasticity reflects a usage externality, whereby one call leads to another[40].

Barnett and Kaserman (1998) caution about the limits of the network externality hypothesis as a justification for subscriber subsidies. They make three important points. First, network externalities are mostly inframarginal at high penetration levels, and it is unnecessary to subsidize the bulk of subscribers who would join the network anyway. Second, economic efficiency is increased by targeting subscriber subsidies at marginal consumers who are most likely to generate network externalities. For example, these might be individuals who receive more calls than they make, and do not value communication sufficiently to subscribe without a subsidy. Third, subscriber subsidies only improve welfare if the external benefits of subscription from the network externality exceed the efficiency losses from financing the subsidies. These arguments lead the authors to the bottom-line conclusion that uniform subsidies are unlikely to improve average consumer welfare.

Although this conclusion is probably overdrawn, Barnett and Kaserman's three cautions are well taken. In particular, it is clearly desirable to target universal service support more efficiently. Third degree price discrimination, which offers discounts to selected consumer groups, or second degree price discrimination based on optional calling plans, are ways to do this.

3.4. Third degree price discrimination

Notwithstanding the attractive properties of Ramsey rules, a simple two-part tariff is not the best way to achieve universal service goals. The efficiency burden of maintaining universal service can be lessened by allowing price discrimination. Economists distinguish various kinds of price discrimination. First-degree price discrimination is charging different prices to different people based on their

[39] Implicit in this discussion is the idea that it may be possible to draw inferences about network externalities from changes in usage prices. The economic consequences of disconnecting someone from a network is not much different from charging an exceedingly high price for telephone calls. It may be possible to draw an inference about network externalities by extrapolating the consequences of small price change.

[40] This usage externality is discussed also by Taylor (2001).

identity. Leaving aside the question of its legality, an effective first degree price discrimination scheme is infeasible for mortal regulators because it requires an omniscient knowledge of consumers' preferences. Second-degree price discrimination is something of a misnomer, because all consumers are offered the same menu of choices and elect different items on the menu according to their preferences. Thus consumers end up paying different prices under second degree price discrimination because they choose to do so. Third-degree price discrimination charges different prices to groups of consumers based on observable characteristics of the group. Different prices based on income or location are examples.

Third degree price discrimination is a recognized tool for promoting universal service. The FCC's low-income and high-cost support policies, discussed in more detail in the next section, fall into this category. Low-income support policies provide discounts to individuals meeting certain means tests. High-cost support policies seek to narrow price differences based on the average cost of service in different locations.

The analytics of optimal third degree price discrimination are a straightforward generalization of the normative theories presented earlier. Suppose consumers are divided into two classes, Class I and Class II, and consider the theory of optimal two-part tariffs with access scale economies but no network externalities (a further generalization to allow for network externalities is pretty straightforward). In general, the two classes may have different demand characteristics and different costs of service. The Ramsey formulas for optimal usage and access prices generalize readily, with notation analogous to before. For Class I, the prices are

$$\frac{p_{\mathrm{I}} - c_{\mathrm{I}}}{p_{\mathrm{I}}} = \frac{\lambda}{1+\lambda}[1 - \varpi_{\mathrm{I}}]\frac{1}{\epsilon_{\mathrm{I}}}$$

$$\frac{r_{\mathrm{I}} - m_{\mathrm{I}}}{r_{\mathrm{I}}} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_{\mathrm{I}}}$$

and for Class II

$$\frac{p_{\mathrm{II}} - c_{\mathrm{II}}}{p_{\mathrm{II}}} = \frac{\lambda}{1+\lambda}[1 - \varpi_{\mathrm{II}}]\frac{1}{\epsilon_{\mathrm{II}}}$$

$$\frac{r_{\mathrm{II}} - m_{\mathrm{II}}}{r_{\mathrm{II}}} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_{\mathrm{II}}}$$

The optimal pricing policies for the two classes are linked by a common value of the Lagrange multiplier λ, which captures the social cost of meeting the expected profit constraint. The linkage arises because profits are aggregated across the two consumer classes. Thus, it is possible for profits on one class of consumers to compensate losses on the other.

This theory provides a rationale for low-income support policies. For simplicity, assume that both classes are served jointly and have the same cost of service, or equivalently that costs are "averaged". Assume also that both classes have the same price elasticity of usage, i.e. the two classes have different demand characteristics based only on different distributions of θ. For concreteness, suppose that Class II consumers are more likely to have a greater demand for usage, i.e. a lower value θ (in the sense of first-order stochastic dominance). Given the empirical evidence that usage increases with income (Crandall and Waverman 2000), it is natural to think of Class I as a low income group.

How should universal service support be targeted at low income (Class I) consumers? Applying the simplifying assumptions, the Ramsey formulas imply

$$\frac{p_{\mathrm{I}} - c}{p_{\mathrm{I}}} = \frac{[1 - \varpi_{\mathrm{I}}]}{[1 - \varpi_{\mathrm{II}}]} \frac{p_{\mathrm{II}} - c}{p_{\mathrm{II}}}$$

$$\frac{r_{\mathrm{I}} - m_{\mathrm{I}}}{r_{\mathrm{I}}} = \frac{\eta_{\mathrm{II}}}{\eta_{\mathrm{I}}} \frac{r_{\mathrm{II}} - m_{\mathrm{II}}}{r_{\mathrm{II}}}$$

That is, the price-cost markups for the two groups are proportional, although the proportionality factors differ for usage and access. For usage prices, the factor of proportionality depends on the ratios of usage demand for the marginal and average subscribers (ϖ) for each class. If both populations were to face the same prices, then the marginal type would be the same for the two classes, but $\varpi_{\mathrm{I}} > \varpi_{\mathrm{II}}$ because of differing mean values of θ. Thus the proportionality factor for usage prices is less than one, i.e.

$$\frac{[1 - \varpi_{\mathrm{I}}]}{[1 - \varpi_{\mathrm{II}}]} < 1,$$

indicating that Class I consumers should face a lower usage price. For access prices, the proportionality factor is the ratio of access elasticities. Although a common marginal type implies $m_{\mathrm{I}} = m_{\mathrm{II}}$, Class I would have a lower penetration rate because of the less favorable distribution θ, implying $\eta_{\mathrm{I}} > \eta_{\mathrm{II}}$, and indicating that Class I consumers should also get a lower access price. Since, at the point of no price discrimination, optimality conditions for usage and access prices fail in the same direction, it would be desirable both to lower p_{I} (relative to p_{II}) and to lower the price of r_{I} (relative to r_{II}), to bring the proportionality conditions into

balance. This heuristic analysis suggests that optimal low income policies should involve both usage subsidies and access subsidies[41].

The theory of third degree price discrimination also provides a logical basis for high-cost support policies, although the logic is rather different than for low-income support. Suppose that Class I and Class II consumers are identical, except that Class I consumers have a higher cost of access. At the optimum:

$$\frac{p_{\mathrm{I}} - c}{p_{\mathrm{I}}} = \frac{\lambda}{1+\lambda}[1 - \varpi_{\mathrm{I}}]\frac{1}{\epsilon}$$

$$\frac{r_{\mathrm{I}} - m_{\mathrm{I}}}{r_{\mathrm{I}}} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_{\mathrm{I}}};$$

and:

$$\frac{p_{\mathrm{II}} - c}{p_{\mathrm{II}}} = \frac{\lambda}{1+\lambda}[1 - \varpi_{\mathrm{II}}]\frac{1}{\epsilon}$$

$$\frac{r_{\mathrm{II}} - m_{\mathrm{II}}}{r_{\mathrm{II}}} = \frac{\lambda}{1+\lambda}\frac{1}{\eta_{\mathrm{II}}}.$$

There are two interesting possibilities. On the one hand, if the marginal cost of access were the same for both consumer classes, and the difference were entirely in the fixed cost of access, then $m_{\mathrm{I}} = m_{\mathrm{II}}$ implies that both consumer classes should face the same prices. This is the economic logic for "geographic averaging". On the other hand, if the marginal cost of access were greater for Class I, then $m_{\mathrm{I}} > m_{\mathrm{II}}$ implies higher access prices for Class I. The resulting lower penetration rate means that $\eta_{\mathrm{I}} > \eta_{\mathrm{II}}$ and $\varpi_{\mathrm{I}} > \varpi_{\mathrm{II}}$; hence access and usage markups should be lower for Class I. Thus, some degree of geographic price discrimination is efficient when marginal access costs vary locationally. The price differences between the two classes for access and usage should move in opposite directions, even though the markup differences move in the same direction.

The fact that geographic price discrimination sometimes is efficient does not imply that the two geographic regions should be priced separately based on their respective costs. If the two classes were treated independently, then Class I would necessarily have higher markups to cover its higher access cost. Consequently, the structure of prices would be the same for both classes, except $\lambda_{\mathrm{I}} > \lambda_{\mathrm{II}}$. This means that it would be economically efficient to relax the profit constraint on Class I customers, and to tighten the constraint on Class II customers to make up

[41] This theoretical analysis has not been developed much in the literature on optimal pricing. It is worth much more attention.

the difference. This could be accomplished by balanced subsidies and taxes on the firms serving Class I and Class II consumers. These transfers should proceed until $\lambda_{\mathrm{I}} = \lambda_{\mathrm{II}}$ resulting in the optimal structure. Service to Class I consumers should operate at a deficit, recovered from profits (or taxes) on service to Class II consumers. This is almost a stylized description of federal high-cost policies in the United States. The difference is that in practice high income areas do not receive a usage subsidy, and perhaps receive an excessive access subsidy.

3.5. Second degree price discrimination

Optional tariffs are an example of second-degree price discrimination. Consumers are offered a choice of service plans, and allowed to self-select the plan that is best. In particular, consumers could be offered a range of service plans that trade off the access price against the usage price. Low volume consumers would prefer a plan with a lower access price and a higher usage price, and conversely for higher volume consumers. The optimal menu of service plans can be constructed using what are now well accepted methods from the mechanism design literature in economics.

The following analysis sketches the mechanism design approach to constructing an optimal menu of service plans, and characterizes the price distortions embedded in those plans. Let $[p(\theta), r(\theta)]$ denote the service plan chosen by a type θ consumer. Ignoring network externalities, the consumer enjoys a consumer surplus of

$$V(\theta) = \theta u(p(\theta)) - r(\theta)$$

Using standard analytical tools (i.e. the envelope theorem and integration), it can be shown that consumers maximize utility by choosing from the menu so that

$$V(\theta) = \int_{\theta_o}^{\theta} u(p(s))ds,$$

and that average consumer surplus over the entire population is

$$\bar{V} = \int_{\theta_o}^{\infty} u(p(\theta))[1 - F(\theta)]d\theta.$$

Now consider profits. Sales to a type θ consumer are

$$X(\theta) = \theta x(p(\theta)) \equiv -\theta u'(p(\theta))$$

and access revenues are related to usage prices according to

$$r(\theta) = \theta u(p(\theta)) - V(\theta).$$

Allowing for scale economies, the profit earned on the type θ consumer is

$$\pi(\theta) = [\theta u(p(\theta)) - V(\theta)] + [p(\theta) - c]\theta x(p(\theta)) - \bar{h}(\theta_o),$$

and average population profit is

$$\bar{\Pi} = \int_{\theta_o}^{\infty} \{\theta u(p(\theta)) + [p(\theta) - c]\theta x(p(\theta)) - \bar{h}(\theta_o)\} f(\theta) d\theta - \bar{V}.$$

Maximizing the Lagrangian $L = \bar{V} + (1 + \lambda)\bar{\Pi}$ with respect to this price function yields the modified Ramsey formula

$$\frac{p(\theta) - c}{p(\theta)} = \frac{\lambda}{1 + \lambda} \frac{1}{\varepsilon} \left[\frac{1 - F(\theta)}{f(\theta)\theta} \right].$$

This formula depends on the hazard rate $\frac{f(\theta)}{1-F(\theta)}$, which is the probability of being a type θ consumer conditional on not being a lower type. If the hazard rate is increasing in θ, as it is for many common distributions, and the average profit constraint is binding ($\lambda > 0$), as it is in the presence of scale economies, then the usage mark-up is smaller for higher volume users[42]. For higher volume users, the usage price is closer to marginal cost. The access price is correspondingly higher for higher volume users, i.e. $r'(\theta) = -\theta x(p(\theta))p'(\theta) > 0$. Moreover, since usage is priced above marginal cost, it is immediate from the break-even constraint that $r(\theta) < \bar{h}(\theta_o)$ for at least some users. An optimal menu of service plans results in higher volume users selecting a plan with a lower usage price and higher access price. The usage price optimally is set above marginal cost for all but the highest volume users, and the access price is below the average cost of access for lower volume users[43].

Cain and MacDonald (1991) provide some econometric evidence supporting the desirability of optional tariffs for local service. Their demand estimates show that, if a measured service option is available for local service, then telephone penetration is insensitive to the monthly charge for flat rate service. This result is consistent with the idea that marginal consumers opt for measured service when

[42] This generalizes the formula for an unregulated monopolist. See Tirole (1988 p. 156).

[43] Faulhaber and Panzar (1977) is an early analysis of the issue. Riordan (2000) considers the $c = 0$ case and shows that a choice of two extreme service plans is optimal. High volume users would choose a flat rate plan with unlimited long distance usage. Low volume users would choose a cheaper plan with prohibitively expensive long distance usage. By continuity, an extreme two-option menu is approximately optimal for c positive but sufficiently small. As a practical matter, the marginal cost of usage is dropping with technological advance and rapidly approaching zero.

given the choice. Cain and MacDonald interpret their results in the following way (1991 p. 303):

> These estimates suggest that universal service can be maintained and expanded, even while more of the NTS financial burden is shifted to local charges. In particular, since telephone subscribership is sensitive to measured access charges, universal service goals can be met, at relatively low cost, by introducing and expanding budget measured service options.

Riordan (2000) points out that similar principles can be applied to long distance usage. In particular, consumers (or long distance companies acting as their agents) can be offered optional access arrangements, or, equivalently, optional arrangements for contributing to a universal service fund. Offered the choice, higher volume users would select a higher fixed monthly payment and lower usage-sensitive payment. Such an arrangement would better target universal service subsidies to marginal consumers.

4. Positive economics of universal service

4.1. Cross-subsidies in the price structure?

Commentators frequently decry cross-subsidies in the structure of telecommunications prices. The AT&T divestiture was based partly on a claim of cross-subsidies running from local to long distance services (Temin 1990). In contrast, the frequent claims today are that business cross-subsidizes residential, long distance subsidizes local, and urban subsidizes residential services. While the term "cross-subsidy" is used loosely even in the academic literature, economists typically are complaining that some set of services (residential, local, or rural) is priced below its long run incremental cost (LRIC). This appears to have become the "popular" meaning of cross-subsidy.

Twenty-five years ago, Faulhaber (1975) sought to discipline the discussion of cross-subsidies by advancing a formal definition and corresponding tests. He defined a subsidy-free price structure as one whose revenues do not exceed the stand-alone cost for any subset of services[44]. Moreover, assuming weak economies of scope, subsidy-free prices must also cover the incremental cost of any subset of services[45]. The stand-alone and incremental cost tests are equivalent for a zero-profit firm. If the firm makes positive economic profits, then cross-subsidies are indicated by a failure of the stand-alone test applied to the whole product set, even though no product need fail the incremental cost test. Thus, the popular meaning of a cross-subsidy in a regulated price structure is justified in Faulhaber's (1975) framework if the firm is held to zero economic profits.

[44] The stand-alone cost is the cost of producing the relevant services in isolation.

[45] The incremental cost is the cost-saving from not producing these services. The necessary and sufficient condition for the equivalence result is that the services are produced subject to weak economies of scope.

Temin (1990) recognizes Faulhaber (1975) by defining a "cross-subsidizing service" as one priced above stand-alone cost, but still accepts popular usage by defining a "cross-subsidized service" as one priced below LRIC. If the firm were to earn positive economic profits, then, by this terminology, it would be possible in the presence of joint costs to have a service receiving a cross-subsidy, but no other service doing the cross-subsidization. Temin meant these definitions to apply only to environments in which rate of return regulation held total profits to zero, e.g. the old Bell system[46]. In this case, a failure of incremental cost test for some group of services, necessarily implies a failure of the stand-alone test for other services.

A possible tension between the popular meaning and Faulhaber's definition of a cross-subsidy is revealed in the following quotation from Kaserman and Mayo (1994 pp. 135–6):

> To some extent, the argument over whether a subsidy exists is semantic. The answer hinges upon one's definition of a subsidy and how one would measure the costs of the services involved. Regardless of the position one adopts, however, there is no economic justification for a system that places the burden of fixed network costs on usage-sensitive prices. Such a system is inefficient whether or not a subsidy results. Consequently, one need not become mired in the subsidy debate to make definite statements about efficient pricing policies. We will continue to use the cross-subsidization terminology throughout the remainder of this article because it is convenient to characterize the overpricing of one service along with the underpricing of another as a cross-subsidy, whether or not these prices fall outside the range that the Faulhaber criteria define. What is more, we are convinced that such cross-subsidization exists, is substantial, and is an accurate description of the existing price structure in this industry.

Kaserman and Mayo's blanket condemnation of price distortions implicitly denies the importance of scale economies and network externalities. As discussed earlier, normative theory provides a rationale for recovering fixed network costs from usage sensitive prices under these conditions. However, more importantly for the discussion at hand is Kaserman and Mayo's insistence on evaluating the merits of price structures in terms of economic efficiency. This is undoubtedly the principal perspective of economists when discussing cross-subsidy issues. Economists' complaints about cross-subsidies typically are on normative grounds: prices below LRIC encourage an overexpansion of telecommunications networks and are a barrier to more efficient entrants.

In contrast, Faulhaber (1975) had a more practical preoccupation. He was concerned that prices above stand-alone cost were not sustainable in a competitive market. The reason is that an equally efficient entrant could successfully undercut a price above stand-alone cost. This is an important issue for universal service, especially in the wake of the 1996 Telecommunications Act. The Act intends to open all telecommunications markets to competition. To the extent that universal service implicitly is supported by Faulhaber cross-subsidies, these subsidies are likely to be undermined by new competition. Recognizing this, the Act

[46] Personal communication with the author.

requires that implicit subsidies be made explicit and portable[47]. State regulators have been concerned about too much competition until new universal service mechanisms are in place. So far, there has been substantial new entry into business markets and not much entry into residential markets, suggesting cross-subsidies flowing from business to residential services. The existence of such a business-to-residential cross-subsidy has been established empirically by Palmer (1992). Rosston and Wimmer (2000b) estimate that nationally the average revenue per line for local service is $39.14 for business lines compared to $18.29 for residential.

A problem with the stand-alone test is that the stand-alone cost of a group of services typically is not observed and therefore is difficult to estimate (Curien 1991). Palmer (1992) addressed this issue for the case of two services by deriving an upper bound on the stand-alone cost under a non-decreasing returns to scale assumption. Using this bound Palmer derived a pair of sufficient conditions for prices to satisfy the stand-alone and incremental cost tests for subsidy-free prices. Palmer estimated costs and revenues for 32 suburban central offices operated by New England Telephone in the mid-to-late 1980s. Almost all of these central offices failed the stand-alone test and a majority failed the incremental cost test. On average, residential revenue fell short of the lower bound on incremental cost by $0.39 per line per month, implying a business-to-residential subsidy of at least $3.45 per business line. These results suggest a substantial business-to-residential subsidy. However, Palmer does not provide confidence intervals or otherwise address estimation errors.

There is some controversy and confusion in the literature about whether long distance services cross-subsidize local services. The stylized fact is that the revenues from local services do not recover their stand-alone costs while the revenues from toll services exceed their incremental costs. The following statement by Curien (1991 p. 91) is typical:

> In telecommunications industries all over the world, the local networks run a deficit, i.e. the connection and subscription charges which are paid by users for their access fail to recover the cost of building and maintaining the connection line and other non-traffic sensitive equipment. As a result, the non-traffic-sensitive costs are subsidized by the revenues derived from traffic and especially from trunk traffic.

Such an assertion apparently flies in the face of Faulhaber's (1975) definition of a cross-subsidy. Indeed, the conditions identified by Curien satisfy Faulhaber's

[47] A portable subsidy is paid to whichever firm provides services. The flip side of the sustainability argument is that services priced below their stand-alone costs are immune to new competition from equally efficient entrants. This appears to be the case for residential local access services in rural areas. Thus, these areas should not expect much local competition unless there is a portable explicit subsidy that makes up the difference. The FCC has recently established limited portable subsidies for the highest cost wire centers in the highest cost states, but largely has left to the states the problem of creating local competition in high-cost rural areas. See Rosston and Wimmer (2000).

conditions for subsidy-free prices[48]: the price of access is below its stand-alone cost, and the price of usage is above its incremental cost. Gabel (1995) builds on this point, arguing that the access services provided by the local loop should be interpreted as a shared input into local exchange and toll services. The published literature does not contain any rigorous showing of a cross-subsidy from toll services to local exchange services[49].

It is also widely held that geographic averaging results in a cross-subsidy from urban to rural services. This follows almost immediately for a zero profit firm under the reasonable assumptions that the stand-alone cost of urban service is substantially less than the stand-alone cost of rural services, and that joint costs are small. However, if the firm is making significant positive profits, then the validity of the claim is less clear. In the United States, regulated local exchange carriers are allowed to earn positive profits on unregulated vertical services, e.g. voice mail and call forwarding. The published literature lacks a rigorous demonstration of an urban-to-rural cross-subsidy that takes account of the profits from vertical services.

4.2. Low income subsidies

In the United States, universal service subsidies are targeted at low-income households via the Lifeline (LL) and Link-Up (LU) programs established by the FCC at the end of 1984. The LL program reduces the monthly cost of telephone service of eligible low income households by an amount equal to $7.00 currently[50]. States provide additional support resulting in total monthly subsidies typically ranging between $5.25 and $10.50; the Virgin Islands is an anomaly with total support of $14.05. The LU program subsidizes the installation charges of a new subscription for eligible households up to $30 plus up to $200 in interest on deferred payments. Eligibility criteria for both programs are established by the individual states subject to FCC approval and vary widely (Federal Communications Commission 1999). Together, the federal components of these programs are projected to cost $480 million in 1999 (Eisner 2000).

Schement, Belinfante and Povich (1997 pp. 193–6) identify twelve states who experienced large increases in telephone penetration for low income households between 1984 and 1994: Connecticut, Georgia, Hawaii, Michigan, Nevada, New Mexico, North Carolina, South Carolina, Tennessee, Vermont, Washington, and Wyoming. Two-thirds of these states were among the early adopters of the federal

[48] Curien's (1991 p. 91) characterization of a "cross-subsidy from traffic to access" is based on an *ad hoc* approach of using "revenue trade-offs" to measure cross-subsidies. The revenue trade-off approach arbitrarily allocates profits and costs to services, including joint and common costs, and asks whether service revenues recover allocated costs plus profits.

[49] See L. Taylor (1993), W. Taylor (1993), Kahn (1993), Gabel and Kennet (1993), and Gabel (1995) for debate on whether access should be regarded to be an input or a separate service.

[50] This is twice the federal subcriber line charge (SLC). The SLC is scheduled to increase to $5.00 under a recent FCC access reform order. Presumably, the LL subsidy will increase commensurately.

low-income support programs. This casual evidence suggests that LL and LU programs have been effective at promoting universal service.

There is also some more rigorous empirical evidence showing that low-income subsidies have increased telephone penetration rates, although the quantitative impact appears to be small relative to the cost of these programs[51]. Table 4 reports selected regressions from three different studies: Garbacz and Thompson (1997; hereafter G&T); Eriksson, Kaserman and Mayo (1990; hereafter EKM); and Crandall and Waverman (2000; hereafter C&W). The three studies employ different data; G&T examines state-level data from the 1990 census; EKM examines annual state-level data from the Current Population Survey; and C&W examines 1990 census data at the level of town. The three studies also employ different specifications, and report the significance of estimates differently[52].

G&T estimate a logit model of state-level penetration, and conclude that the LL and LU programs have a statistically significant but small marginal effect on penetration for the average state. Their explanatory variables include the monthly price of (flat rate) local service, and the installation charge for new accounts. Demographic variables include the percent of households living below the poverty line, and the percent of households living in urban areas. The key variable for testing the effectiveness of low income subsidies is the amount of LL and LU funds paid out per poor household in the state. Although G&T interpret their regression equation as a demand equation, the price variables are not significant[53].

EKM report a related analysis based on pooled state-level cross section and time series data for the period from 1985 through 1993 and draw similar conclusions. The annual penetration data is drawn from the Current Population Survey, which Garbacz and Thompson (1997; 2000) criticize as being more subject to measurement error than the decennial census data, resulting in unreliable estimates. Also worrisome is that EKM apparently ignore serial correlation in the error terms for each state, which could bias their statistical tests. EKM find a positive significant effect of LL and LU subsidies only in states that have a large poor population[54].

[51] Park and Mitchell (1989) show in a calibrated simulation model that Lifeline rates are unlikely to significantly increase penetration.

[52] See also Albery (1995) for a related study.

[53] This could be due to endogeneity bias. Prices of local service and installation are regulated by the states. The coefficients on these variables would be biased toward zero if states with low penetration rates tended to choose lower prices for residential service. (The LL and LU estimates could suffer similar endogeneity bias; see the discussion of C&W below.) G&T do find significant price coefficients in other specifications.

[54] EKM include 1984 penetration in all of their specifications as an explanatory variable "in order to standardize for the cross-sectional variation in the observed penetrations rates prior to the sample time period." It is unlikely that the relationship is stable over time; why should penetration levels in 1993 and 1998 bear the same relation to 1984? It is not clear *a priori* how this source of specification error might bias the estimated effects of the low income subsidies. G&T show in their study that inclusion of lagged penetration does not much matter.

Table 4
Effectiveness of low income subsidies

Study data source	G&T[55] 1990 census	EKM[56] 1985–93 CPS	C&W[57] 1990 census
Dependent variable (test statistic)	$\ln \frac{penetration}{1-penetration}$ (standard error)	*penetration* (t-statistic)	*penetration* (t-statistic)
Constant	3.35*	0.54622*	1.003*
	(0.728)	(16.879)	(146.6)
Local service price	0.009	−0.00103*	0.00017
	(0.008)	(4.103)	(0.94)
Installation charge	−0.003	−0.00032*	−0.00070*
	(0.003)	(3.824)	(9.51)
Long distance price		−0.10593*	0.00096
		(6.064)	(0.44)
p.c. income			0.00048*
			(7.09)
% poor	−8.757*	−0.00200*	−0.282*
	(0.728)	(7.041)	(6.43)
% poor squared			0.292*
			(2.83)
% urban	0.473*		
	(0.132)		
% rural		−0.00013	
		(1.628)	
Population density			0.0047*
			(5.22)
% black		−0.00040*	−0.034*
		(4.0060)	(5.30)
% Hispanic		−0.00039*	−0.032
		(2.926)	(5.82)
penetration in 1984		0.50301*	
		(16.036)	
p.h. LL-LU subsidy < O		−0.00605*	
		(2.142)	
p.h. LL-LU subsidy × % poor		0.00059*	
		(2.482)	

[55]This is regression (2) in Table 4 of Garbacz and Thompson (1997).
[56]Model A in Table 2 of Eriksson, Kaserman, and Mayo (1998).
[57]Model (1a) in Table 5-5 of Crandall and Waverman (2000).

Table 4 (*continued*)

Study data source	G&T[55] 1990 census	EKM[56] 1985–93 CPS	C&W[57] 1990 census
Dependent variable (test statistic)	$\ln \frac{penetration}{1-penetration}$ (standard error)	*penetration* (t-statistic)	*penetration* (t-statistic)
p.h. LL-LU subsidy ÷ % poor	0.017*		
	(0.002)		
LL dummy × % poor			0.016
			(1.30)
LU dummy × % poor			−0.098*
			(5.06)
p.h. high cost payments		−0.00009	
		(0.413)	
R^2	–	.8424	0.736
# observations	44	432	1,97

*Statistically significant at 0.05 level.

Both G&T and EKM interpret the estimated quantitative significance of the low income subsidies with the aid of "policy experiments". G&T estimate from their regression analysis that an across the board 10% increase in subsidies would increase average penetration by "substantially less than one tenth of one percent." EKM conclude that an additional $10,000 in subsidies would add only 18 new subscribers for a state whose poverty level is average, and 75 new subscribers for the poorest states. While these calculations are provocative, the policy interpretations are not really valid, because the parameter estimates on which they are based do not have clear structural interpretations. In particular, the models do not distinguish whether the increased subsidy levels of the policy experiment come from more generous support levels or more generous eligibility criteria.

To illustrate how eligibility criteria might matter consider the following simple model. Suppose that a subsidy of s dollars is targeted at households below the poverty line, but that the prevailing eligibility criterion results in only a fraction λ of poor households being able to receive the subsidy. Suppose further that households above the poverty rate choose to have a telephone with probability β_1, subsidized poor households with probability β_2, and unsubsidized poor households with probability β_3, with $\beta_1 > \beta_2 > \beta_3$. If POV is the poverty rate, then the observed penetration rate would be

$$PEN = \beta_1(1 - POV) + \beta_2\lambda POV + \beta_3(1 - \lambda)POV$$

and the subsidy per household would be

$$SUB = s\lambda POV.$$

Thus, looser eligibility criteria (i.e. higher λ) increases both the penetration rate and the amount of subsidy. Solving these two equations to eliminate λ gives

$$PEN = \beta_1 - (\beta_1 - \beta_3)POV + \left(\frac{\beta_2 - \beta_3}{s}\right)SUB.$$

Therefore, holding constant the amount of the subsidy (s), the penetration rate is decreasing in the poverty rate and increasing in the subsidy per household (SUB). In this specification, the subsidy per household is serving as a proxy for eligibility criteria. This simple model provides some justification for including per household subsidies directly into a penetration equation, but also suggests that functional form may be important and that the parameter estimates need to be interpreted carefully. In this example, a doubling of subsidy payments corresponds to the policy experiment of doubling the size of the eligible population. The effect of this experiment on measured penetration would be $\beta_2 - \beta_3$. Thus, the estimated coefficient on SUB would have to be multiplied by s to measure the effect of the policy change on telephone penetration.

A priori, C&W seems the most interesting of the three studies because it relies on more disaggregated data. The study matches price data to census data on towns (cities, or designated places). The price data were obtained directly from large local exchange carriers, resulting in 1896 observations. The study measures the effect of LL subsidies with a dummy variable for the state's implementation of the program interacted with the poverty rate. Effectively, this is measuring whether poor communities in states who have LL programs in place have higher penetration rates than similar poor communities in states lacking LL programs. The regression analysis does not find a significant effect of LL on the measured penetration rate. This seems consistent with their related finding that the effect of local service prices is not significant either. These results suggest that LL has not been an effective policy tool for advancing universal service. It is possible that the supporting estimates suffer from endogeneity bias, although this seems less likely than in G&T and EKM, because in C&W the regulated prices and subsidy policies are set at the state level while penetration is measured at the town level.

C&W measure the effect of LU simply as a dummy variable interacted with poverty, effectively comparing penetration rates of poor towns in states with and without the LinkUp program. The regression equation finds that the LinkUp policies have a statistically significant *negative* effect on telephone penetration. This paradoxical result seems hard to explain, and appears inconsistent with the

finding that higher installation charges reduce penetration. C&W suggest that the result is due to the fact that only two states, Delaware and Illinois, lacked LU programs and that the regulators in these states declined to implement LU because penetration rates were already high. In other words, the estimated coefficient suffers from an endogeneity bias. In view of this potential problem, the C&W study does not appear to provide very convincing evidence on the effectiveness of Link-Up.

4.3. High cost subsidies

Telephone companies serving high-cost areas in the U.S. receive direct subsidies. Federal subsidies to companies serving high-cost areas have been paid out under a variety of mechanisms (Federal Communications Commission 1999). "High-cost loop support" has been given to companies with above average non-traffic sensitive costs. Additional "long term support" subsidizes a uniform below-cost carrier line rate for participating companies. Finally, "local switching support" defrays some of the traffic sensitive costs of companies serving small market areas. Taken together, these mechanisms provided $1.7 billion in assistance in 1999. A new high-cost program established in 2000 consolidated the subsidies to larger companies in a new cost fund, and established intrastate subsidies based on forward-looking economic cost and targeted to high-cost wire centers within the receiving state. The Telecommunications Act requires that implicit universal subsidies be made explicit and financed by taxes ("contributions") on the revenues of telecommunications companies. The federal programs are financed by taxes on interstate and international revenues.

Eriksson, Kaserman and Mayo (1998) studied the effectiveness of high-cost support on the prices of Bell Operating Companies (BOCs) with the following regression equation

$$PRI = 15.53250 + 0.014660 \cdot CST - 20.20702 \cdot BUS - 0.13469 \cdot USF$$

where PRI is a weighted average flat rate for residential service, CST is the historical cost of "outside plant" for providing local access in the rate base, BUS is the ratio of business and residential lines, and USF is high cost support per household paid from the Universal Service Fund. These variables are measured at the state level. Although the coefficients are all statistically significant, the R^2 of this regression equation is only 0.20. The regression indicates a negative correlation between the amount of high cost support and the price of local service. This estimated equation suggests that an extra dollar of high-cost support translates into only a 13 cent reduction in the price of local service.

Thus, given a low price elasticity for local access, this suggests that high-cost subsidies paid to companies are not very effective at increasing penetration rates. Indeed, Eriksson, Kaserman and Mayo (1998 p. 498) conclude that a $10,000 increase in BOC high-cost support would add only 15 subscribers at a cost of $666 per new subscriber. As above, this "policy experiment" is suggestive, but not definitive because the estimated parameters lack clear structural interpretations.

Recent FCC policy has left the problem of high-cost support largely to the state jurisdictions. Rosston and Wimmer (2000a) ask what level of state universal service funds would be necessary to cover the forward-looking economic costs of local service under the assumption that telephone companies earn $32 per line, which is a benchmark revenue level that the FCC had considered previously as relevant for establishing high-cost support levels. They estimate that the state high-cost subsidies would come to almost $3 billion in the aggregate, the financing of which would require consumers to pay a weighted-average tax rate of 2.41% on intrastate revenues. They further estimate that, if instead of establishing high-cost subsidies, the states rebalanced rates to reflect costs, then telephone penetration rates would drop by only one-half of one percent nationwide. This calculation leads them to question whether this modest effect on penetration is worth the efficiency loss created by the distortionary revenue taxes, and to recommend that high-cost support be targeted better to low-income households.

5. Conclusions

A number of conclusions can be drawn from this survey of issues about universal residential telephone service. First, the two important "underserved" populations in the United States are the poor and Native Americans. These populations have substantially lower residential telephone penetration rates even after controlling for locational, demographic, and cost factors. Second, although penetration rates for similar communities are different in different parts of the United States, differences in state regulatory policies account for no more than 1–2% of this variation. Third, the extent to which "taxes" on long distance usage are an inefficient means of public finance for universal service programs depends on details of the industrial organization of long distance telephone services. Fourth, while scale economies and especially network externalities provide potentially important theoretical rationales for universal service policies, the empirical evidence on their quantitative significance is scant and inconclusive. Fifth, optional tariffs governing local and long distance toll services potentially are effective devices for targeting implicit subsidies for local access. Sixth, there is some econometric support for the proposition that business

rates have cross-subsidized residential rates, according to the formal economic definition of a cross-subsidy, but the frequent claims that long distance cross-subsidizes local and that urban cross-subsidizes rural services rest on more casual appraisals. Seventh, although economic theory provides rationales for well-designed low-income and high-cost support policies for promoting universal service, the limited empirical evidence on the issue suggests that low income and high-cost subsidies have at best a quantitatively small impact on penetration rates relative to their cost.

The main conclusion of the chapter, though, is that there remains a shortfall of research on the economics of universal service. First, the determinants of telephone penetration are still not completely understood. For example, it is unclear why Native American populations suffer lower telephone penetration even after controlling for poverty, climate, and costs. It is also unclear to what extent price regulation and universal service policies explain state-specific variations in telephone penetration. Second, the empirical importance of scale economies and network externalities as rationales for universal service remains cloudy. For example, more information on usage profits earned by service providers on marginal subscribers would permit a better calculation of the economic opportunity cost of expanding basic access services. A serious attempt to estimate the quantitative significance of "long distance network externalities" from price elasticities for long distance services would contribute usefully to the policy debate. Evidence on the significance of offsetting call externalities is also sorely needed. Third, an empirical quantification of the potential welfare gains from implementing optional tariffs, or other forms of second-degree price discrimination, seems to be within reach of modern structural econometrics with a sufficiently rich data set (Miravete 2000). Fourth, well-crafted tests of the propositions that long distance has cross-subsidized local services and that urban have cross-subsidized rural services are long overdue. Fifth, a fully convincing appraisal of the performance of low-income and high-cost programs in advancing universal service awaits better data and more careful econometrics. Settling these issues for the paradigm problem of maintaining and advancing basic universal residential telephone service will strengthen the foundations for debating and evaluating the next generation of universal service policies.

Only a few qualified lessons can be drawn for policy-makers. First, while state regulators should "benchmark" their regulatory and universal service policies to other states, the adoption of "best practices" might increase residential telephone penetration by only a few percent. Second, even though policy-makers can in good faith remain hesitant to embrace too closely the chorus of calls for strict cost-based pricing of local access services, the economic case for a significant markup of usage prices is debatable. Third, while the FCC and the states should consider optional arrangements for universal service contributions as a better way to target universal service support, the quantitative significance of such

policies remains an open question. Fourth, the FCC most likely should exempt service provided to Lifeline and LinkUp recipients from universal service contributions. All such advice is tentative, of course, pending further economic research.

Although beyond the scope of this chapter, it is worth mentioning, in closing, a few upcoming issues. One new issue is universal service auctions. The 1996 Telecommunications Act opens the door for the FCC to consider auctions as an alternative mechanism for high-cost support. The FCC has so far refrained from doing so, although in its 1997 *Universal Service Order* expressed an intention to open a proceeding on the matter. In the mid 1990s, California considered but did not adopt auctions for awarding state high-cost support. Other places, including Europe and Australia, have also considered auction mechanisms for high cost support. There is a new theoretical literature on the topic (Laffont and Tirole 2000; Sorana 2000). Another new issue for which there is an emerging literature is the effect of universal service policies on competition (Gasmi, Laffont, and Sharkey 2000; Choné, Flochel, and Perrot 2000). The Telecommunications Act requires that universal service policies in the United States be competitively neutral. In the U.S. and even more blatantly in other countries, new competitors pay taxes to incumbents to help finance the incumbents' universal service obligations. Armstrong (2001a, 2001b) argues that a well-designed universal service policy, together with cost-based access pricing, nevertheless can provide efficient incentives for entry and make-or-buy decisions. A third emerging issue is a broader definition of universal service, discussed by Crandall and Waverman (2000). There is considerable and growing political pressure to further expand the definition of universal service to encompass Internet access. Downes and Greenstein (1999) show empirically that access to Internet services is already widely available, albeit at very different speeds in different places. Cremer (2000) develops a theoretical argument that network externalities might be particularly strong for broadband Internet service. These are all likely to be among the important universal service policy issues in the coming decade.

6. Appendix: Variable definitions and summary statistics for Table 1

6.1. Census data

The following variables were created from the 1990 Census STF-3 files. Each variable is measured at the Census Block Group (CBG) level.

Penetration is the fraction of occupied housing units in the CBG with a telephone in the housing unit.

%*Poor* is the fraction of CBG population living below the poverty line.

Median income is the median household income of the CBG, measured in thousands of dollars.
%*Female head of household* is the fraction of households in the CBG with a female head of household.
%*Senior* the fraction of CBG population that is 65 years of age or older.
%*Children* the fraction of CBG population that is 15 years of age or younger.
%*High school* is the fraction of CBG population with a high school degree, including those with some college but no college degree.
%*College* is the fraction of CBG population with a college degree.
%*Black* is the fraction of CBG population that is black.
%*Hispanic* is the fraction of CBG population that is of Hispanic origin. If a person is black, white, Asian, etc., and also of Hispanic origin, then they are counted only as being Hispanic.
%*Native* the fraction of CBG population that is Native American.
%*Asian* the fraction of CBG population that is Asian.
%*Other nonwhite* the fraction of CBG population that is nonwhite and not a member of the aforementioned race categories.
Population density is the number of people, measured in thousands, per square kilometer living in the CBG.
Wire center population is the number of people, measured in thousands, living in the area serviced by the same wire center that services the CBG. This variable was created from the 1990 Census STF-3 files, but only after linking the CBGs to wire centers using data obtained from the FCC.

6.2. Climate data

In order to measure the effect of climate on telephone penetration, data from the United States Historical Climatology Network (U.S. HCN) was linked to the census data[58]. The U.S. HCN data is measured at the station level, identified by its latitude and longitude. Each CBG was assigned to the station with the minimum product of absolute differences between latitude and longitude. Data is available from 1221 stations for the 48 contiguous state, although data from Tennessee was missing. Data for Alaska, Hawaii, and the District of Columbia

[58] The U.S. HCN data is made publicly available by the Carbon Dioxide Information Analysis Center. For more information see: Easterling, D. R., T. R. Karl, E. H. Mason, P. Y. Hughes, D. P. Bowman, and R. C. Daniels, T. A. Boden (eds.). "United States Historical Climatology Network (U.S. HCN) Monthly Temperature and Precipitation Data." ORNL/CDIAC-87, NDP-019/R3. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory. Oak Ridge, Tennessee, 1996.

are not available from this source. A fully quadratic form was specified for the following variables:

Temperature is the annual mean temperature in 1989 recorded by the station, within state, nearest to the CBG.

Precipitation is the total precipitation in inches in 1989 recorded by the station, within state, nearest to the CBG.

6.3. Cost data

The FCC has published an economic-engineering model that estimates, among other things, the forward-looking economic cost of providing basic local service[59]. This model incorporates locational data and 1996 quantity and price data into an optimization model. The cost estimation procedure is based on the FCC's TELRIC (total element long run incremental cost) methodology. The CBGs are matched to wire centers. Given the relatively small increase in telephone penetration rates in recent years, the relative forward-looking costs probably have not changed too much between 1990 and 1996, except that boundaries of wire centers do change occasionally. For given wire center assignments locational data, e.g. terrain, which are a critical determinant of cost differences, certainly remain constant.

Not every CBG can be matched to a wire center. The model uses a selection of wire centers in Bellcore's LERG database. Only wire centers which were listed as end offices, hosts or remotes, and which were not owned by wireless, long distance or competitive access providers were used. This left roughly some 12,000 wire centers, covering roughly half of the original sample of CBGs. When wire centers are matched to the CBGs for which weather data is available, roughly forty percent of the original sample of CBGs were left.

The cost variables used in the estimation are defined as follows.

Loop length is an estimate of the average length of the connection of the customer to the wire center, including distribution (the cable connecting a customer to a Serving Area Interface (SAI)) and feeder (the cable connecting an SAI to a wire center) distances.

Average forward looking cost is the FCC's estimate of the average monthly forward-looking cost of providing basic local service, including distribution, feeder and end-office switching costs, measured in dollars.

[59] See Sharkey (2001) for a description of the FCC's Hybrid Proxy Cost Model.

6.4. Summary statistics

Table 5
Summary statistics

	(I)		(II)		(III)–(V)	
	mean	s.d.	mean	s.d.	mean	s.d.
Penetration	93.9	9.0	94.4	7.9	94.3	7.8
% Poor	14.0	14.1	12.8	12.6	12.7	12.5
Median income	31.2	16.4	31.9	15.9	31.7	15.7
% Female h.o.h.	11.8	10.4	10.6	8.9	10.7	9.0
% Senior	13.3	9.2	12.9	9.0	13.1	8.8
% Children	23.8	9.2	23.7	9.0	23.6	8.8
% High School	31.8	9.5	32.1	8.9	32.2	8.7
% College	16.3	12.3	17.0	12.2	16.7	12.0
%Black	12.4	25.1	10.1	21.2	10.9	22.0
% Hispanic	7.6	16.5	7.4	15.8	5.7	13.6
% Native	0.9	4.9	0.7	3.8	0.7	3.4
% Asian	2.2	6.3	2.4	6.5	1.9	5.0
% Other nonwhite	0.1	0.6	0.1	0.5	0.1	0.5
Pop. density	2.9	4.8	1.9	5.0	2.0	5.4
W.c. population			29.7	26.3	28.3	25.7
Loop length					21.0	19.0
Average f.l. cost					31.4	25.2
Temperature					53.8	8.2
Precipitation					42.0	16.0
# Observations	222,264		116,715		95,171	

References

Albery, B., 1995, What level of dialtone constitutes "universal service"? Telecommunications Policy, 19, 365–80.

Armstrong, M., 2001a, Access pricing, bypass, and universal service, American Economic Review, Papers and Proceedings, 81, 297–301.

Armstrong, M., 2001b, The theory of access and interconnection pricing, in M. Cave, S. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Amsterdam: Elsevier Science, BV.

Barnett, A.H. and D. L. Kaserman, 1998, The simple welfare economics of network externalities and the uneasy case for subscribers subsidies, Journal of Regulatory Economics, 13, 245–54.

Brown, S. J. and D. S. Sibley, 1986, The theory of public utility pricing, Cambridge: Cambridge University Press.

Cain, P. and J. M. MacDonald, 1991, Telephone pricing structures: The effects on universal service, Journal of Regulatory Economics, 3, 293–308.

Choné, P., L. Flochel and A. Perrot, 2000, Universal service obligations and competition, Information Economics and Policy, 12, 249–59.

Crandall, R. W. and L. Waverman, 2000, Who pays for "universal service"? When telephone subsidies become transparent, Washington D.C.: Brookings Institution Press.
Cremer, J., 2000, Network externalities and universal service obligation in the Internet, European Economic Review, 44, 1021–31.
Curien, N., 1991, The theory and measure of cross-subsidies, International Journal of Industrial Organization, 9, 73–108.
Downes, T. and S. Greenstein, 1999, Do commercial ISPs provide universal access?, in S. Gillet and I. Vogelsang, eds., Competition, Regulation, and Convergence: Current Trends in Telecommunications Policy Research, Mahwah, N.J.: Lawrence Erlbaum Associates.
Dyer, P., 1997, Households without telephones in the UK, Telecommunications Policy, 4, 341–53.
Eisner, J., 2000, State-by-state telephone and universal service data, Industry Analysis Division, Common Carrier Bureau, Federal Communications Commission, January.
Eriksson, R. C., D. L. Kaserman and J. W. Mayo, 1998, Targeted and untargeted subsidy schemes: Evidence from post-divestiture efforts to promote universal service, Journal of Law and Economics, 41, 477–502.
Faulhaber, G. R., 1975, Cross-subsidization: Pricing in public enterprises, American Economic Review, 65, 966–77.
Faulhaber, G. R. and J. C. Panzar, 1977, Optimal two-part tariffs with self-selection, Bell Laboratories Economic Discussion Paper, No. 74.
Federal Communications Commission, 1999, Monitoring Report, CC Docket No. 98-202, December.
Federal Communications Commission, 2000, Report and Order, CC Docket No. 96-45, May.
Gabel, D., 1995, Pricing voice telephone services: Who is subsidizing whom?, Telecommunications Policy, 19, 453–64.
Gabel, D. and D. M. Kennet, 1993a, Pricing of telecommunications services, Review of Industrial Organization, 8, 15–20.
Gabel, D. and D. M. Kennet, 1993b, Pricing of telecommunications services: A reply to comments, Review of Industrial Organization, 8, 43–48.
Gasmi, F., J.-J. Laffont and W. W. Sharkey, 2000, Competition, universal service and telecommunications policy in developing countries, Information Economics and Policy, 12, 221–48.
Garbacz, C. and H. G. Thompson, Jr., 1997, Assessing the impact of FCC Lifeline and Link-Up programs on telephone penetration, Journal of Regulatory Economics, 11, 67–78.
Garbacz, C. and H. G. Thompson, Jr., 2000, FCC telephone penetration rate data: A caveat, Telecommunications Policy, 24, 969–79.
Greene, W. H., 1993, Econometric analysis, Upper Saddle River, N.J.:, Prentice Hall, 2nd edition.
Hausman, J., 1998, Taxation by telecommunications regulation: The economics of the E-rate, Washington, D.C.: American Enterprise Institute Press.
Hausman, J., 1999, Economic welfare and telecommunications regulation, Yale Journal of Regulation, 16, 19–51.
Hausman, J., 2002, Mobile telephone, in M. Cave, S. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science, B. V., 563–604.
Hausman, J., T. Tardiff and A. Belinfante, 1993, The effects of the breakup of AT&T on telephone penetration in the United States, American Economic Review, Papers and Proceedings, 83, 178–84.
Hudson, H., 1995, Access to telecommunications in the developing world: Ten years after the Maitland report, in G. Brock, ed., Toward a Competitive Telecommunications Industry: Selected Papers from the 1994 Telecommunications Policy Research Conference, Mahwah, N.J.: Lawrence Erlbaum Associates, 235–50.
International Telecommunications Union, 1999, Yearbook of Statistics: Telecommunications Chronological Time Series, 1988–1997, ITU, January.
Katz, M. L. and C. Shapiro, 1994, System competition and network effects, Journal of Economic Perspectives, 8, 93–115.

Kahn, A. E., 1993, Pricing of telecommunications services: A comment, Review of Industrial Organization, 8, 39–42.
Kaserman, D. L. and J. W. Mayo, 1994, Cross-subsidies in telecommunications: Roadblocks on the road to more intelligent pricing, Yale Journal on Regulation, 11, 119–147.
Kaserman, D. L. and J. W. Mayo, 2001, Competition in the long distance market, in M. Cave, S. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Amsterdam: Elsevier Science, BV.
Kaserman, D. L., J. W. Mayo and J. E. Flynn, 1990, Cross-subsidization in telecommunications: Beyond the universal service fairy tale, Journal of Regulatory Economics, 2, 231–49.
Laffont, J.-J. and J. Tirole, 2000, Competition in Telecommunications, Cambridge, MA.: MIT Press.
Liebowitz, S. J. and S. Margolis, 2001, Network effects, in M. Cave, S. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Amsterdam: Elsevier Science, BV.
Maher, M. E., 1999, Access costs and entry in the local telecommunications network: A case for de-averaged rates, International Journal of Industrial Organization, 17, 593–609.
Mankiw, G. N. and M. D. Whinston, 1986, Free entry and social inefficiency , RAND Journal of Economics, 17, 48–58.
Miravete, E., 2000, Quantity discounts for taste-varying consumers, CARESS working paper, University of Pennsylvania.
Mitchell, B. M., 1978, Optimal pricing of local telephone service, American Economic Review, 68, 517–37.
Mueller, M. L., Jr., 1997, Universal service: Competition, interconnection, and monopoly in the making of the American telephone system, Cambridge, MA: MIT Press, and Washington, D.C.: AEI Press.
Mueller, M. L., Jr. and J. R. Schement, 1996, Universal service from the bottom up: A study of telephone penetration in Camden, New Jersey, The Information Society, 12, 273–93.
Palmer, K., 1992, A test for cross subsidies in local telephone rates: Do business customers subsidize residential customers?, RAND Journal of Economics, 415–31.
Park, R. E. and B. M. Mitchell, 1989, Local telephone pricing and universal service, RAND Corporation Report R-3724-NSF, June.
Prieger, J., 1998, Universal service and the Telecommunications Act of 1996, Telecommunications Policy, 22, 57–71.
Riordan, M. H., 2000, An economist's perspective on residential universal service, in I. Vogelsang and B. M. Compaine, eds., The Internet Upheaval: Raising Questions, Seeking Answers in Communications Policy, Cambridge, MA: MIT Press.
Rosston, G. L. and B. S. Wimmer, 2000a, The ''state'' of universal service, Information Economics and Policy, 12, 261–83.
Rosston, G. L. and B. S. Wimmer, 2000b, From C to shining C: Competition and cross-subsidy in telecommunications, in S. Greenstein and B. Compaine, eds., 2000 TPRC Proceeding .
Sappington, D., 2001, Price regulation, in M. Cave, S. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Amsterdam: Elsevier Science, BV.
Schement, J. R., 1995, Beyond universal service: Characteristics of Americans without telephones, 1980–1993, Telecommunications Policy, 2, 477–85.
Schement, J. R., A. Belinfante and L. Povich, 1997, Trends in telephone penetration in the United States 1984–1994, in E. M. Noam and A. J. Wolfson, eds., Globalism and Localism in Telecommunications, Amsterdam: Elsevier, 167–199.
Schmalensee, R. 1981, Monopolistic two part tariff arrangements, Bell Journal of Economics, 12, 445–56.
Sharkey, W., 2002, Representation of technology and production, in M. Cave, S. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science, B. V., 179–222.

Sorana, V., 2000, Auctions for universal service subsidies, Journal of Regulatory Economics, 18, 33–58.

Taylor, L. D., 1994, Telecommunications Demand in Theory and Practice, Dordrecht, Netherlands, and Boston: Kluwer Academic Publishers.

Taylor, L. D., 1993, Pricing of telecommunications services: Comment on Gabel and Kennet, Review of Industrial Organization, 8, 15–20.

Taylor, L. D., 2002, Customer demand analysis, in M. Cave, S. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science, B. V., 97–142.

Taylor, W. E., 1993, Efficient pricing of telecommunications services: The state of the debate, Review of Industrial Organization, 8, 21–38.

Temin, P., 1990, Cross subsidies in the telephone network, Journal of Regulatory Economics, 2, 349–62.

Tirole, J., 1988, The theory of industrial organization, Cambridge, MA: MIT Press.

Vogelsang, I. and B. Mitchell, 1991, Telecommunications pricing: Theory and practice, Cambridge: Cambridge University Press.

This Page Intentionally Left Blank

SECTION III – COMPETITION

This Page Intentionally Left Blank

Chapter 11

COMPETITION POLICY IN TELECOMMUNICATIONS

DANIEL F. SPULBER*

Northwestern University

Contents

* I acknowledge helpful comments from Shane Greenstein, Sumit K. Majumdar, and Ingo Vogelsang. The chapter draws some material from my article Spulber (1995).

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

Over the course of the 20th century, antitrust policy in the United States has played a greater role in telecommunications than in any other industry[1]. Although the effects of competition, technological change, and federal and state regulation have been significant and well recognised, antitrust interventions altered the development of the telecommunications industry at many critical junctures. In this chapter, I consider the principal ways that antitrust doctrines have been interpreted and applied in the telecommunications industry, without attempting to give a full review of the extensive scope of federal telecommunications antitrust. In light of the Telecommunications Act of 1996, I briefly consider the prospects for antitrust policy in the 21st century.

Competition policy in telecommunications has consistently held to three major themes: monopolisation, interconnection, and mergers. The three major government antitrust actions addressed monopolisation and leveraging of market power through foreclosure and tying activities. Private antitrust actions addressed predatory pricing, connection of terminal equipment, and interconnection to other telecommunications service providers. Although the issue of collusion has received some attention in the economics literature, it does not seem to have played much of a role in competition policy in the telecom sector.

The Department of Justice (DOJ), the Federal Trade Commission (FTC), the Congress, and the Federal Communications Commission (FCC), with input from state regulatory commissions and state attorneys general, formulate competition policy in telecommunications. Those telecommunications companies that are regulated utilities are partially exempt from antitrust with authority and jurisdiction ceded to the FCC and state public utilities commissions (Waldman, 1986). Extensive antitrust enforcement activities of the Department of Justice gradually removed the cover provided by regulation leading to the eventual break-up of the Bell system monopoly.

The history of telecommunications antitrust spans almost a century with sporadic bursts of intense legal and regulatory activity. History matters in antitrust because of the great importance of case law interpretations of the basic antitrust statutes. A good part of the history of competition policy in the U.S. telecommunications industry is the history of a single company: AT&T. Early on, regulatory decisions and competition policy set the stage for AT&T's virtual monopoly over telephony. Incorporated in 1885 and restructured almost a century later in 1984, AT&T was the subject of numerous private antitrust actions and three major government antitrust cases. Over fifty private antitrust suits were filed against AT&T regarding the company's actions between 1968 and 1982[2].

[1] For a discussion of competition policy in other countries see for example Vogelsang (1997).
[2] For a useful overview of the private antitrust suits against AT&T see Philips (1988).

Even before the expiration of the original patents of Alexander Graham Bell in 1893 and 1894, the Bell system faced public criticism for monopolisation (Grosvernor and Wesson, 1997). The Bell system followed a policy of refusing to connect to independents. In 1907, John Pierpont Morgan formally assumed control of AT&T and assisted in its policy of undercutting or acquiring independents. By 1913, AT&T faced threats by the Wilson administration to break up or even nationalise the company. Upon the death of J. P. Morgan that year, the Bell System's Theodore Vail offered to divest the company's interest in the Western Union telegraph company and to connect the Bell System to independents and share long-distance revenues with them[3].

The chapter is organised as follows. Part 2 concerns monopolisation and examines predatory pricing, raising rivals' costs, natural monopoly and market power in telecommunications. Part 3 addresses leveraging and looks at essential facilities, barriers to entry, tying, and cross subsidisation. Part 4 discusses mergers since the Telecommunications Act of 1996. Part 5 concludes the discussion.

2. Monopolisation

In popular discussions about network industries, there is often confusion between three distinct concepts: monopolisation, monopoly and natural monopoly. *Monopolisation* refers to company behaviour directed at achieving a monopoly, whether or not the company is successful in that attempt. *Monopoly* refers to the market position of a company having no established competitors, although there may be potential entrants. *Natural monopoly* refers to technology and market demand conditions: there is said to be a natural monopoly if a single company can supply the amount demanded by the market at less cost than two or more firms.

Antitrust policy targets monopolisation behaviour only. In its Standard Oil decision, the Supreme Court introduced the Rule of Reason by identifying as a violation of the Sherman Act unreasonable attempts to monopolise, not the condition of monopoly itself. Telecommunications providers such as AT&T often had a monopoly by virtue of their status as regulated utilities, although such regulation did not exempt them from antitrust review of anticompetitive behaviour. Because antitrust policy is concerned with monopolisation behaviour, it also does not identify natural monopoly technology as a violation of the Sherman Act. The courts and antitrust policy makers raised the question of whether telecommunications technology has exhibited or continues to exhibit natural monopoly properties. Such technology in itself is not a sufficient basis for antitrust scrutiny but was alleged to be an instrument of monopolisation.

[3] Grosvernor and Wesson (1997, pp. 167, and 240–242).

2.1. Monopolisation

Section 2 of the 1890 Sherman Act forbids monopolisation or attempts to monopolise. Three major federal antitrust suits that challenged American Telephone and Telegraph (hereafter AT&T) illustrate the application of policy restrictions on monopolisation. All three actions ended in settlements. These suits and some private antitrust actions are listed in Table 1[4].

In a 1913 suit, the federal government charged that AT&T monopolised long distance telephone service between Oregon and Washington and between Washington and Idaho by interconnecting only with its own local affiliates rather

Table 1
Selected public and private bell system cases

Case	Year initiated/ completed	Issue	Outcome
U.S. v. AT&T	1913/1914	Monopolisation of long distance	Consent decree, Kingsbury commitment
U.S. v. Western Electric	1949/1956	Monopolisation of equipment manufacturing	Consent decree, Final Judgement
U.S. v. AT&T 524 F. Supp	1974/1982	Monopolisation of long distance and equipment manufacturing	Consent decree, Modified Final Judgement
Hush-a-Phone Corp. v. U.S. 238 F. 2d 266 (D.C. Cir. 1956)	1956	Attachment to customer premises equipment	D. C. Circuit Court reverses FCC and allows competitor to sell attachments
Carter v. AT&T 365 F. 2d. 486, Carterfone 13 F. C.C.2d	1966/1967	Right to connect devices to network	FCC allows connection to network of customer premises equipment not made by Bell
MCI v. AT&T MCI 708 F. 2d 1081	1974/1986	Cross subsidisation, predatory pricing, refusal to interconnect	Damages paid to MCI
Southern Pacific Communications Co. V. AT&T 566 F. Supp. 825	1978/1984	Cross subsidisation, predatory pricing, refusal to interconnect	D.C. Circuit court of appeals affirms lower court decision accepting AT&T's defence

[4] Kellogg et al. (1992, pp. 137–248).

than other companies[5]. The Justice Department and AT&T reached an agreement in 1913 known as the Kingsbury Commitment. Under the settlement, AT&T would not purchase independent phone companies without prior approval by the Interstate Commerce Commission and would allow independents to interconnect with its system[6]. Shortly thereafter, the Willis-Graham Act of 1921 gave the Interstate Commerce Commission authority to approve telephone company mergers, providing AT&T with exemptions from antitrust restrictions[7].

In 1949, the federal government charged AT&T with monopolising telecommunications equipment and services through its patent licensing policies. The case was settled with a 1956 consent decree referred to as the Final Judgement in which AT&T agreed to license Bell Labs patents and to restrict manufacturing by its Western Electric subsidiary to telecommunications equipment. The decree restricted AT&T to regulated natural monopoly services[8].

Finally, in 1974, the federal government charged AT&T with monopolisation of long distance services and equipment manufacturing. The case was settled in a 1982 consent decree known as the Modification of Final Judgement (MFJ)[9]. The decree eliminated the prior 1956 consent decree. The break-up of the Bell System, which took place on January 1, 1984, represented a large-scale vertical divestiture that changed the structure of the industry. AT&T retained its long-distance business, equipment manufacturing including Western Electric, and its Bell Labs research unit. AT&T was required to divest its local operating companies. The local exchange services were divided among seven Regional Bell Operating Companies (RBOCs): Ameritech, Bell Atlantic, BellSouth, Nynex, Pacific Telesis, Southwestern Bell, and U S West.

In turn, the RBOCs were quarantined in the local exchange markets by three line-of-business restrictions. The MFJ forbade the RBOCs from providing long-distance services from one local access and transport area (LATA) to another. The RBOCs were prevented from manufacturing telecommunications equipment[10]. The third restriction concerning the provision of information

[5] Kellogg et al. (1992, p. 200).

[6] Horwitz (1989, pp. 100–101)

[7] Willis-Graham Act, 42 Stat. 27, June 10, 1921, see Horwitz (1989, p. 102) for additional discussion. The Willis-Graham Act overrode the Kingsbury Commitment, see Kellogg et al. (1992, p. 201). Kellogg et al. observe that between 1921 and the Communications Act of 1934, the ICC approved 272 out of 275 proposed acquisitions of independents.

[8] Kellogg et al. (1992, pp. 200).

[9] United States v. American Tel. & Tel. Co., 552 F. Supp. p. 131 (D.D.C. 1982), aff'd sub nom. Maryland v. United States, 460 U.S. p. 1001 (1983).

[10] The local exchange network is the portion of the public switched network served by the local exchange carriers which include the seven RBOCs as well as hundreds of independent telephone companies. The designated areas served by the RBOCs are referred to as LATAs. The RBOCs provide both exchange services (e.g. basic dial tone service, call waiting, call forwarding, Centrex) and exchange access services (e.g. connection to long distance carriers). The interLATA or itnerexchange network is the portion of the public switched network that is served by the long-distance carriers. See North American Telecommunications Association (1991, pp. 19–22).

services was effectively removed after divestiture[11]. Moreover, the MFJ required the RBOCs to provide equal access to the local network to all long distance carriers.

The MFJ effectively placed the Federal District Court (District of Columbia) and the Department of Justice in the regulation business. Peter Huber (1997) observes that antitrust law created a new *ad hoc* commission: "This is nothing to be very proud of, for commission-by-decree is antitrust law at its worst." Moreover, the FCC was drawn further into the industry as a promoter of competition, particularly through administration of open access. As I have noted elsewhere, this constituted a major role reversal in which antitrust enforcement took on the role of utility regulator while federal regulators assumed the task of making competition policy (Spulber, 1989).

The role reversal is unfortunate in both regards. Judges are not well suited as industry regulators; they lack the technical expertise and administrative support personnel to keep up with industry developments and end up being a bottleneck, as was the case with Judge Greene's long control of the telecommunications industry under the MFJ. Although the judiciary is best suited for adjudicating property, tort or contract disputes, it lacks the rule making and investigative capacity of regulatory agencies. The DOJ also is not suited as an industry regulator in comparison with administrative agencies. The DOJ as a general purpose agency lacks industry expertise and specialised rulemaking capacity and is therefore not suited to long-term monitoring of a specific industry. The DOJ is much less subject to external influence from consumers and industry groups, thus averting the public notice and comment that is the bedrock of American administrative law. Finally, the DOJ has a somewhat different relationship with Executive and Legislative branches of government thus altering the essential oversight of administrative agencies.

Conversely, the FCC and state regulators are less well suited to promotion of competition. Regulatory agencies should be impartial rule makers and interpreters of relevant legislation. As promoters of competition, regulators may be tempted to pick technology winners and attempt to manage the competitive process rather than setting general rules that do not favour individual companies or technologies. The FCC and state regulators have a limited role as advisors to the antitrust agencies in merger policy as concerns companies in the telecommunications sector, including establishing preconditions on merger approval.

The Telecommunications Act of 1996 (hereafter the 1996 Act), which superseded the MFJ, redressed this *topsy turvy* administrative situation by increasing the regulatory oversight of the FCC and ending the control of

[11] For thorough discussions of the divestiture and its aftermath, see Kellogg et al. (1992), Crandall (1991) and Temin (1987).

Judge Greene's court. However, the involvement of the DOJ as regulator continues in a different form. Rather than participating in enforcement of the MFJ through the courts, the DOJ acts as an advisor to regulators at the FCC. Under the Act, the FCC gives substantial weight to DOJ recommendations.

The DOJ continues its regulatory oversight of the telecommunications industry particularly through the checklist provisions of the 1996 Act. The checklist is a highly technical set of conditions requiring nondiscriminatory access to the local exchange for competitive local exchange companies and a showing of facilities-based competition in the local exchange. The 1996 Act required open access to the local exchange under the checklist as a condition for allowing the RBOCs to enter into long distance. Thus, the 1996 Act continued the inter-LATA line-of-business restrictions on the RBOCs by its checklist provisions. The DOJ has intervened repeatedly in telecommunications markets through the checklist process. For example, in November, 1999 the DOJ found that Bell Atlantic had complied with most of the checklist provisions and recommended that the FCC approve the company's petition subject to conditions. In February 2000 the DOJ recommended that the FCC reject the application of SBC Communications to enter long distance in Texas. The assistant attorney general of the Antitrust Division stated that "The department found that SBC has not shown that it is providing nondiscriminatory access to its local lines"[12]. The FCC subsequently approved SBC's entry into long distance in Texas.

2.2. Predatory pricing and raising rivals' costs

The government's case that led to the MFJ charged AT&T with various types of anticompetitive behaviour. The government's characterisation of AT&T's pricing practices were consistent with the notion of predatory pricing[13]. Accusations of predatory pricing, cross subsidisation and refusals to interconnect also arose in MCI v. AT&T and in Southern Pacific Communications v. AT&T. Moreover, the government charged AT&T with practices that raised rivals' costs.

Predatory pricing is said to exist if firms incur a loss with the intention of eliminating rivals and the objective of later raising prices to recoup earnings after the rivals have exited from the market[14]. The Supreme Court has found, based on legal and economic analysis, that predatory pricing is unlikely to succeed because the incumbent firms must incur losses with little guarantee of

[12] Schiesel (2000, p. C8).

[13] Private antitrust suits alleged that AT&T had engaged in predatory pricing. In *Northeastern*, the 2nd Circuit Court applied Areeda and Turner's marginal cost standard as a test for predatory pricing, see Northeastern Telephone Co. v. American Telephone and Telegraph Co., 1980–81 Trade Cases, Par 64, 027 pp. 76, 315–76, 316 (2d. Cir 1981), cited in Phillips (1988, p. 722). On the marginal cost standard see Areeda and Turner (1975).

[14] See, e.g., Baumol and Sidak (1994, p. 63) and Spulber (1989, pp. 475–476).

successful recoupment, rivals can also incur losses in anticipation of future profits, and new entrants will appear if prices are subsequently raised[15]. In establishing predatory pricing, it is difficult to distinguish in practice between low competitive prices and predatory prices or to distinguish low earnings from so-called predatory losses[16].

The difficulty of establishing predatory pricing in practice and the inconsistency of the predation framework with economic incentives is reflected in the nature of the government's case. The government charged that AT&T had priced "without regard to cost," not only because the company lacked the necessary cost information but also because the government alleged that the pricing practices were intended to deter entry into long distance. This argument had two significant antitrust implications. First, it effectively removed the cost standard typically used in analyses of predatory pricing since the government did not use cost data as the basis for its allegation. Thus, the government alleged predation without demonstrating that the company priced below cost. Second, the argument placed antitrust responsibilities on regulated firms, even though the prices offered by such firms are approved by regulators (Noll and Owen, 1989).

Raising rivals' costs allegedly creates monopoly power by providing a price umbrella so that a company can raise prices and undercut its competitors (Salop and Scheffman, 1983) and (Krattenmaker and Salop, 1986). An incumbent firm able to raise rivals' costs could potentially deter entry by raising the costs of potential entrants. Krattenmaker and Salop (1986) argue that a company could raise costs of competitors by purchasing exclusionary rights from input suppliers, outbidding rivals for inputs because of the monopoly rents that exclusion of rivals would yield. Such an argument is predicated on the supply of inputs being price inelastic and the notion that an incumbent firm can corner the market on critical inputs profitably. However, the prices on inputs would be expected to increase due to competitive bidding by rival firms in its own industry and other input users outside the industry[17].

[15] Brooke Group Ltd. V. Brown & Williamson Tobacco Corp., 113 S. Ct. 2578 (1993); Matsushita Elec. Indus. Co. v. Zenith Radio Corp., 475 U.S. 574, 589 (1986).

[16] The law and economics literature has substantially refuted the notion of predatory pricing, see particularly McGee (1958; 1980) and Bork (1978, pp. 144–155). See Spulber (1989) for further discussion. Researchers in industrial organization have applied game-theoretic models in which behavior resembling predatory pricing emerges in equilibrium, see Ordover and Saloner (1989) and Lott (1999) for surveys of industrial organization models of predation. Game theoretic models of predation emphasize the role of asymmetric information and reputation building. Lott (1999) presents an empirically based critique of these game-theoretic models of predation by considering managerial compensation and entrenchment in firms that have been accused or convicted of predation.

[17] Evidence supporting the notion that a company can raise the costs of its rivals is problematic. Lopatka and Godek (1992) show that Alcoa had limited control over inputs and that exclusionary agreements were ancillary to its purchasing agreements. Alcoa purchased a very small fraction of electric power production and the company's purchases of Bauxite had little effect on its price or availability.

The government argued that AT&T used the regulatory process to increase the costs of competitors. By obtaining regulatory prohibitions on the connection of customer premises equipment (CPE) to the telephone system, or requiring the use of unnecessary protective attachments, AT&T raised the cost to providers of competitive CPE[18]. It was further alleged that AT&T abused the regulatory process by excessive regulatory and judicial actions that raised the legal cost of competitors and thus deterred their entry[19]. Such arguments seem circular. Presumably, utility regulation is intended to control competitive entry, utility prices and customer service. If the regulatory process itself is subject to abuse as an anticompetitive instrument by the incumbent utility, modification of the regulatory process would appear to be a more direct remedy than supplementing imperfect regulation with antitrust activity. If competitive entry is desirable, this suggests that there is little need to continue traditional utility regulation. In that case, the direct route to avoiding alleged abuse of the regulatory system as an anticompetitive instrument would be to dismantle the regulatory system itself, as some have suggested[20].

The feasibility of a telecommunications company raising rivals' costs depends on whether a firm could successfully control access to a critical supplier or critical inputs without rivals discovering alternative suppliers or substitute inputs. Even if a firm can increase the price of critical inputs by bidding competitively for those inputs, its increased costs would be likely to outweigh any competitive benefits. As with predatory behaviour generally, the returns to raising rivals' costs after rivals are driven from the market must be greater than the costs of driving them out. Moreover, once rivals are driven out, the incumbent firm must be able to defend the market against future entrants to secure returns to predation. Moreover, a company does not have any incentive to raise its rivals' costs in a competitive market. For example, AT&T divested its Bell Labs and Western Electric manufacturing unit to form Lucent so that the spin-off could realise greater sales of equipment to RBOCs and other potential competitors of AT&T. There would have been little gain to AT&T from raising the price of switching equipment it was attempting to sell to competitors who could then turn to alternative suppliers[21].

[18] Noll and Owen (1989, pp. 402–403).

[19] Noll and Owen (1989, p. 404).

[20] Huber (1997, p. 201) argues for abolishing the FCC entirely in favor of common law and antitrust law: "Get rid of the Commission and the constitutional debates go away, too. But so does the Commission's power to anesthetize the antitrust laws. Those laws, all in all, protect competition far better than a commission-tolerant Constitution."

[21] The notion that an incumbent firm could raise the costs of its rivals through discriminatory access charges to its network is similar to the leverage problem to be addressed further in the next section.

2.3. Natural monopoly

Under Theodore Vail, with his well-known slogan "One policy, one system, and universal service," AT&T maintained that telephone service was a natural monopoly (Horwitz, 1989). This view was reflected in the Communications Act of 1934 which instituted federal regulation (Breyer, 1982). In the AT&T case, the DOJ contended that the "natural monopoly characteristics" of the local exchange precluded competition "until a point of concentration of interexchange traffic above the end office"[22]. According to the Court of Appeals, the premise of the antitrust case against AT&T and the Decree was that "AT&T had used its natural monopoly over local-exchange services to impede competition in related markets"[23]. In 1987, Judge Greene stated: "The exchange monopoly of the Regional Companies has continued because it is a natural monopoly"[24]. Thus, while focusing on monopolisation behaviour, the DOJ and the courts viewed natural monopoly as a potential anticompetitive instrument.

A given production technology is said to exhibit the property of *natural monopoly* if a single firm can supply the market at lower cost than can two or more firms[25]. If the technology of local exchange telecommunications exhibits natural monopoly, then a single firm could presumably construct and operate that network at a lower cost than can two or more firms[26]. A sufficient

[22] Response of the United States to Comments Received on the BOC LATA Proposals, pp. 9–10, United States v. Western Elec. Co., No. 82-0192 (D.D.C. Nov. 23, 1982).

[23] United States v. Western Elec. Co., 894 F.2d 1387, 1389 (D.C. Cir. 1990).

[24] United States v. Western Elec. Co., 673 F. Supp. pp. 525, 537–38 (D.D.C. 1987). "Exchange telecommunications," he continued, "is characterized by very substantial economies of scale and scope." Id. P. 538, Judge Green relied on the arguments of others. The quoted statement references the Senate testimony of then Attorney General William Baxter, Hearing before the Senate Committee on Science and Transportation, 97th Cong., 2d Sess. 59.

[25] The concept of natural monopoly is generally credited to John Stuart Mill (1848, chapter 9). Mill emphasizes the problem of wasteful duplication of transmission facilities that can occur in utility services. The connection between natural monopoly and regulation is developed by Leon Walras (1936) with reference to the construction and operation of railroads.

[26] Statistical studies showing that costs are not subadditive include Shin and Ying (1992) who find that the costs of local exchange carriers were not subadditive before the AT&T divestitutre using data from 1976 to 1983, and Evans and Heckman (1984) who show that AT&T's costs were not subadditive. Estimating telecommunications network costs is problematic for companies that have been regulated because data are obtained from regulatory accounting information. Also, the data are often presented at a aggregate level that is not suited to evaluation of cost functions. The estimation of cost functions using standard econometric techniques is difficult at best since an established legacy system built up over decades is not likely to be optimized. Engineering cost models that make assumptions about system configurations need not describe the costs of existing systems. Moreover, the notion of comparing the costs of two identical systems serving the same geographic area is likely to be counterfactual. See the chapters by Fuss and Waverman and Sharkey in this *Handbook*.

condition for a single-product cost function to have the natural-monopoly property is for the technology to exhibit economies of scale over the relevant range of output[27]. Multiproduct cost functions exhibit natural monopoly if and only if the cost function derived from the underlying technology is subadditive, that is, the cost of producing a set of products is less than any division of production between two firms.

Natural monopoly technology does not preclude competitive entry despite the efficiency of production by a single firm. Even if technology exhibits natural monopoly characteristics, competitors may compete to serve the market (Demsetz, 1968). Moreover, if an incumbent firm prices above cost, entrants can compete for customers. Thus, an industry can still be contestable even if the technology exhibits natural monopoly (Baumol et al., 1982). Even if the incumbent prices at cost there are still some situations in which an incumbent monopoly cannot find prices that sustain its position against entry.

Avoiding duplication of facilities, particularly duplication of the fixed costs of the network system, has been an important component of the natural-monopoly argument for regulation of the local exchange. The argument is that since costs are minimised by not duplicating transmission facilities, regulators should bar the entry of competing carriers. The 1996 Act implicitly assumes that the local exchange is not a natural monopoly because it dismantles regulatory barriers to entry, thus overturning the approach of the Modification of Final Judgement. At the same time, the Act provides for access by mandating interconnection, use of existing network elements and resale of wholesale services, so that the problem of possible duplication is addressed.

Natural monopoly technology is consistent with competition in a more fundamental sense: the technology of entrants differs from that of incumbents. The standard definition of natural monopoly presumes that incumbents and entrants have the same cost function and presumably the same underlying technology[28]. The assumption of a common technology is generally unrealistic, and particularly

[27] Economies of scale are said to be present if the marginal costs of production are less than the average costs of production over the relevant range of output. Economies of scale can be due to many different technological factors, such as specialization of function and division of labor permitted by increased output. Fixed costs are a source of economies of scale that is particularly significant to the telecommunications industry – and to all other industries that require networks, such as railroads, oil and natural gas pipelines, electricity, and water services. Fixed costs are costs that do not vary with fluctuations in output, unlike oeprating or "variable" costs. The fixed costs of establishing a network system are the costs of facilities such as tranmission lines, which are not sensitive to the level of transmission on the lines. The firm's average cost function refers to the cost per unit of output evaluated at each output level. The firm's marginal cost function refers to the additional cost of producing one more unit of output, evaluated at each level of output. Economies of scale at a given output level is not necessary for natural monopoly. The natural monopoly property can be present at an output level at which the cost function exhibits decreasing returns to scale. See Spulber (1989, p. 117) for further discussion.

[28] See for example Baumol et al. (1982, p. 17) in which the definition of natural monopoly refers to an industry in which all of the firms have the same cost function.

so in telecommunications, which has been characterised by rapid technological change. An entrant is unlikely to duplicate the technology of the incumbent's legacy network (Spulber, 1995). Entrants may employ capital equipment of different vintages than the incumbent so that their systems embody different technologies.

Technological change has made available multiple forms of transmission. Even if the traditional copper-wire network used to provide plain old telephone service (POTS) might have been a natural monopoly, there are many other transmission technologies including coaxial cable television systems, fibre-optic cable, various land-based wireless systems, and satellite-based systems. It is difficult to identify a single best technology since each of these transmission approaches has different properties in terms of costs and performance. The many uses of transmission networks, including telephony, mobile communications, data transmission and video, suggest that different transmission technologies are suited to different uses. Entrants that offer specialised networks targeted to particular applications are likely to have a different technology than the incumbent. Moreover, the entrant can target specialised market segments without the need to duplicate the incumbent's system.

Transmission systems involve various mixtures and configurations of these technologies within networks and across interlinked networks. George Calhoun observed that "the abandonment of hierarchical structures is gathering momentum, especially in the core public network and in specialised computer networks." Calhoun (1992) calls it the "laminar network," "a series of partly competing, partly complementary, somewhat differentiated, overlapping access fabrics" that "will consist of multiple layers of transmission facilities for accessing the core network at an increasing number of gateways. The lowest levels will still be copper-based fabrics; the vast installed base of wireline telephony and coaxial cable TV plant that will continue in use for decades. Growing over these there will be several new layers of fibre optic plant – and, because of its nature, ever more layers of digital radio. Even within a given fabric layer there will almost certainly be a great deal of technical diversity."

The advent of Internet-based communication with broadband digital transmission over a fibre-optic backbone, developments in switching and interconnection of multiple networks further expands modes of transmission. Internet transmission entails movement of packets of information rather than dedicated lines of the traditional telephone system. There are also many different means of gaining access to the Internet[29]. Alternative on-ramps to the information superhighway include the traditional telephone system, digital subscriber lines (DSL), the cable television transmission system, fibre optic lines, and wireless access.

Competition from Internet telephony, cable telephony, and wireless providers further increased the alternatives to the traditional telephone system. The natural

[29] For a discussion of pricing and regulation of access alternatives see Sidak and Spulber (1988).

monopoly question was to a great extent rendered moot by extensive entry into the local exchange. Such entry involved transmission facilities that effectively duplicated the transmission facilities of the local exchange incumbents. AT&T's acquisition of Telecommunications Inc. in 1998 and Media One in 1999 were targeted at providing telephone and Internet access over coaxial cable, effectively demonstrating the feasibility of using a second wire to the household for telephony. At the same time, traditional cellular providers faced the entry of digital personal communications service (PCS) providers, adding wireless access for households. Business customers also had DSL, fibre optic and wireless access alternatives to the traditional phone network.

Technological and market developments after the 1996 Act resulted in increased competition of telecommunications alternatives in the local exchange. Competition in long distance communications also increased and was likely to increase further with the entry of the former RBOCs into the market. These developments substantially modified antitrust concern over the effects of natural monopoly on competition in telecommunications.

2.4. Market power

The evaluation of market power in regulated industries is complicated by the fact that regulation restricts prices, constrains entry and imposes various service requirements. Kellogg, Thorne and Huber point out that antitrust involving AT&T helped to establish three important principles regarding antitrust against regulated companies. First, high market shares do not necessarily indicate the presence of market power for a regulated industry. Second, in the presence of regulation, a high market share need not imply that the company has attempted to monopolise the industry. Third, regulatory policies should be taken into account in determining whether refusals to interconnect are an antitrust violation[30].

Market structure in the industry after deregulation inherits features created by regulation. These features adjust at differing rates to competitive forces. In the wake of deregulation under the 1996 Act, the FCC determined that AT&T was no longer a dominant firm in long-distance markets based on their evaluation of AT&T's market power. In an antitrust context, market power is said to arise when a firm is able profitably to raise its price substantially above the competitive level, or to raise market prices by restricting their output[31]. The FCC's Fourth Report and Order cited the Areeda and Turner definition of market power as "the ability to raise prices by restricting output" and the Landes and Posner definition of market

[30] Kellogg et al. (1992, pp. 152–154).

[31] While practically every firm has some market power, due to market frictions such as transaction costs and transportation costs, empirical tests for market power generally specify some threshold level, such as a 5% or 10% increase over the competitive price. The empirical test also must make a determination of the estimated competitive price for the relevant market.

power as "the ability to raise and maintain prices above the competitive level without driving away so many customers as to make the increase unprofitable"[32].

Determination of market power for antitrust purposes includes both demand and supply responses to the firm's price change, including potential entry. A firm's ability to raise its price profitably depends on the extent to which the firm's customers reduce their purchases as a result, which is the own-price elasticity of the firm's demand. The price responsiveness of the firm's demand depends in part on the substitutes available to the firm's customers. In addition, it depends on the reactions of the firm's competitors, as they alter their prices, product offerings, sales and marketing efforts, and amount of services sold. In addition, it is essential to take into account not only the supply response of the firm's existing competitors, but also the supply responses of new entrants that could be attracted to enter by the prospect of higher prices.

To evaluate market power, thus requires a definition of the firm's relevant product and geographic markets. These definitions should be sufficiently broad to identify the demand and supply responses to the firm's pricing decisions. In evaluating AT&T's market power, the FCC defined the product market to include all interstate, domestic, interexchange services. The Commission in its Fourth Report and Order defined a single national market (including Alaska, Hawaii, Puerto Rico, U.S. Virgin Islands, and other U.S. offshore points) as the relevant geographic market[33].

There are several determinative economic reasons for these conclusions. All interstate, domestic, interexchange telecommunications services are the relevant product market because such services generally are sold together. A customer of an interexchange carrier purchases all domestic service to any location as one service, not a la carte. Services are not sold on a point-to-point basis, but rather, a customer is able to call all locations. Thus, such services almost always appear as a bundle, so that unbundling them conceptually would depart from the way the services are both sold and purchased. As a consequence, companies compete by offering complete services and schedules of prices for the overall service. Customers usually choose a single carrier for their entire interexchange service rather than purchasing narrower product services from multiple carriers simultaneously. With the consolidation of RBOCs and the bundling of various types of telecommunications services, the regulatory distinctions between local, regional and long distance service will eventually be erased[34].

The geographic market definition should be national in scope because the major long distance carriers compete nationally offering services that originate at

[32] See the discussion in FCC 95-427, In the Matter of Motion of AT&T Corp. to be Reclassified as a Non-Dominant Carrier, p. 5.

[33] Id. p. 9.

[34] Kellogg et al. (1999) observe "The breakup of Bell in 1984 separated telephone equipment from local service, and local from long-distance. The post-1996 competitors will bring them all back together."

all points and terminate at all points. Competition between these companies takes the form of competition between national networks. Marketing campaigns, sales efforts, pricing programs and promotions also tend to be are national in scope.

In its classification of AT&T as nondominant, the Commission declared that "in the interstate, domestic, interexchange market, supply is sufficiently elastic to constrain AT&T's unilateral pricing decisions"[35]. AT&T, as cited in the Order, stated that in 1993 there were more than 500 carriers providing service in the United States, 394 of which provided equal access service in at least one state, nine carriers provided equal access and service at least 45 states, 81 regional carriers served at least four states, and at least twelve interexchange carriers served every state[36]. As the Commission stated, "AT&T asserts, and no one disputes, that MCI and Sprint alone can absorb overnight as much as fifteen percent of AT&T's total 1993 switched demand at no incremental capacity cost; that within 90 days MCI, Sprint and LDDS/Wiltel, using their existing equipment, could absorb almost one-third of AT&T's total switched capacity; or that within twelve months, AT&T's largest competitors could absorb almost two-thirds of AT&T's total switched traffic for a combined investment of $660 million. The FCC concluded that "AT&T's competitors have sufficient excess capacity to constrain AT&T's pricing policies"[37].

The FCC found further that "residential customers are highly demand-elastic and will switch to or from AT&T in order to obtain price reductions and desired features." Citing AT&T data showing that as many as twenty percent of its residential customers change interexchange carriers at least once a year, the FCC concluded that the "high churn rate among residential consumers – approximately 30 million changes are expected in 1995 – demonstrates that these customers find the services provided by AT&T and its competitors to be very close substitutes"[38]. In addition, the FCC found that business customers also are "highly demand elastic"[39]. They reaffirm their earlier finding that "business users consider the offerings of AT&T's competitors to be similar in quality to AT&T's offerings."

3. Leveraging

Leveraging is a type of monopolisation behaviour that uses monopoly power in one market to obtain market power in the second market. Excluding competitors in the second market is known as *foreclosure*. In *Berkey Photo*, the Second Circuit Court found that a firm violates Section 2 of the Sherman Act "by using its

[35] Id. p. 16.
[36] Id. p. 14.
[37] Id. p. 14–17.
[38] Id. p. 17.
[39] Id. p. 18.

monopoly power in one market to gain a competitive advantage in another, albeit without an attempt to monopolise the second market''[40]. The monopoly leverage argument has been rejected in *Air Passenger Computer Reservation Systems*[41]. The Supreme Court in *Spectrum Sports* suggested that violating the Sherman Act required at least the dangerous probability of monopolisation and the intent to monopolise[42]. The Second Circuit has since affirmed the requirement of harm to competition in the second market[43].

Leveraging has been applied in telecommunications antitrust under both the essential facilities and tying doctrines[44]. The essential facilities doctrine, established in the 1912 *Terminal Railroad* case, targeted anticompetitive behaviour consisting of denial of access to competitors to a facility that would be difficult to duplicate[45]. Under certain conditions, denial of access to essential facilities violates Section 2 of the Sherman Act. The application of the doctrine in the 1973 *Otter Tail* case in the electric utility industry laid the groundwork for its use in telecommunications antitrust[46].

3.1. Essential facilities

The essential facilities doctrine has elements of both natural monopoly and barriers to entry. As in the case of natural monopoly, some policy makers believe that duplication of essential facilities would be inefficient. Also, by virtue of their cost and difficulty of duplication, essential facilities resemble certain forms of barriers to entry. An entrant allegedly cannot duplicate existing facilities economically because, while an incumbent has already incurred the costs, an entrant faces the need to make an irreversible investment with all the attendant economic risks. In addition, it is sometimes argued that the entrant's costs of duplicating the facility are higher than they were for the incumbent.

Antitrust policy is not directed at the possession of essential facilities or the existence or barriers to entry that are due to technology, cost and demand conditions in the industry. Such facilities or entry barriers, if they exist, are not the

[40] Berkey Photo Inc. v. Eastman Kodak Co., 603 F.2d 263, (2d Cir. 1979) p. 276.
[41] Air Passenger Computer Reservation Systems Antitrust Litigation, 694 F. Supp. pp. 1443, 1472 (C.D. Cal. 1988).
[42] Spectrum Sports, Inc. v. McQuillan, 113 U.S. p. 884 (1993).
[43] Twin Laboratories, Inc. v. Weider Health & Fitness, 900 F.2d pp. 566, 570–71 (2d Cir. 1990).
[44] See Kellogg, et al, 1992, pp. 142 and 156–166.
[45] United States v. Terminal Railroad Association (224 U.S., p. 383). However, Reiffen and Kleit (419) show that the case does not support a vertical theory of antitrust harm: ''Consistently misinterpreted, it has served as a source for misbeliefs about the economics of vertical integration.'' In that case a group of railroads were accused of controlling access to bridges over the Mississippi River to St. Louis. Reiffen and Kleit (1990) found instead that the railroads did not exclude anyone and that they charged themselves the same prices that were charged to competitors.
[46] Otter Tail Power Co. v. United States (410 U.S., p. 366). See the discussion in Kellogg et al., 1992, pp. 139–140.

result of anticompetitive behaviour. As with natural monopoly technology, the existence of essential facilities or barriers to entry enters into antitrust policy when used for the purpose of leveraging. The networks of incumbent telecommunications providers were construed to be essential facilities and alleged to provide an instrument for monopolisation of telecommunications services.

The case for the break-up of AT&T argued that the company had a monopoly over the local exchange network. Because the network was alleged to be an essential facility, AT&T was presumed able to use its control over that facility to foreclose competitors in other markets who needed access to the local exchange network to provide services. AT&T was viewed as able to leverage its alleged market power in the local exchange to extend its monopoly to long distance telecommunications, sale of manufactured equipment, and information services.

In 1982, Judge Greene stated with regard to interLATA services: "The complexity of the telecommunications network would make it possible for [the RBOCs] to establish and maintain an access plan that would provide to their own interexchange service more favourable treatment than that granted to other carriers"[47]. Similarly, with respect to the manufacture of telecommunications equipment, Judge Greene stated that "non affiliated manufacturers would be disadvantaged in the sale of such equipment and the development of a competitive market would be frustrated"[48]. The DOJ stated that "[t]he reorganisation of AT&T . . . is intended to eliminate the present incentives of the BOCs . . . to discriminate against AT&T's competitors in the markets for interexchange services, information services, customer premises equipment, and the procurement of equipment used to provide local exchange services"[49].

The essential facilities doctrine arose in MCI Communications Corp. v. AT&T and in Southern Pacific v. AT&T. In *MCI*, the Seventh Circuit found that AT&T's had refused to connect MCI with the local exchange: "a monopolist's control of an essential facility (sometimes called a 'bottleneck') can extend monopoly power from one stage of production to another, and from one market to another"[50].

[47] United States v. American Tel. & Tel. Co., 552 F. Supp. pp. 131, 188 (D.D.C. 1982).

[48] Id. p. 190. With regard to information services, Judge Greene stated: "The Operating Companies would . . . have the same incentives and the same ability to discriminate against competing information service providers that they would have with respect to competing interexchange carriers." Id. p. 189. Ingo Votelsang points out that the two markets are different because competing long-distance providers offer essentially the same product while in ocntrast, information service providers offer a variety of differentiated services that increase the overall market.

[49] Competitive Impact Statement p. 24, United States v. AT&T, No. 74-1698 (D.D.C. Feb. 10, 1982) (footnote omitted). The Decree's injunctive provisions "limit the functions of the divested BOCs to preclude the possibility of a recurrence of the type of monopolizing conduct that the United States alleges to have resulted from AT&T's ownership of regulated local exchange carriers and its simultaneous participation in competitive, or potentially competitive, markets." Id. p. 24.

[50] MCI Communications Corp. v. AT&T, 708 F.2d p. 1132.

Assuming that the local exchange network was an essential facility and that it conferred market power on the incumbent firm, it is questionable whether an incumbent firm would have an economic incentive to engage in leveraging. With fixed output proportions in the two markets, the market power obtained by foreclosing access to the second market would be entirely due to the monopoly in the first market. There would be no need to acquire a monopoly in a second market since monopoly rents could be obtained by raising the cost of access.

Advocates of leveraging modify the argument to apply to regulated companies. If the incumbent firm is subject to rate regulation in some form, its ability to capture monopoly rents in its home market is limited. Accordingly, the firm has an incentive to extend its monopoly to other markets that are not subject to regulation as a means of shifting its monopoly rents beyond the reach of regulators. Rate regulation of local exchange companies who possess a monopoly over the local exchange would induce them to seek monopoly rents in long distance to shift profits out of the local exchange and away from state utility regulators.

The standard antitrust remedy for the essential facilities problem is to mandate equal access[51]. The view that the local exchange is an essential facility is reflected in the MFJ's requiring equal access to the local exchange network for long distance carriers. The 1996 Act took up the equal access mission from the earlier antitrust concern with essential facilities. Enforcing the provisions of the 1996 Act has focused on mandating access to the facilities of the local exchange and pricing of access services. The 1996 Act specified access to the facilities as a *quid pro quo* for the former RBOCs to enter into long distance communications.

The basis of essential facilities doctrine is subject to question. Even if facilities have natural monopoly properties, competitive entry is still feasible as already noted. Moreover, facilities are unlikely to remain essential for long when they can be made obsolete by technological change. The barriers to entry aspect of the essential facilities doctrine is also questionable since technological change facilitates competitive entry. Moreover, contracts between entrants and customers overcome the risks of making 'sunk' investments to compete with the incumbent. Technological advances have the same implications for the essential facilities doctrine even if the incumbent's facilities involve a combination of natural monopoly and sunk costs. Entrants with new technology can realise operating or investment cost savings and offer better performance than the incumbent's system.

3.2. Barriers to entry

In the AT&T case, Judge Greene observed with regards to barriers to entry, "The evidence introduced at the trial of this case clearly demonstrated that duplication of the ubiquitous local exchange networks would require an enormous and

[51] Kellogg et al. (1992, p. 140).

prohibitive capital investment, and no one seriously questions that this is true''[52]. Elsewhere, he stated ''The government alleges that defendants have monopoly power in each of these markets and, to prove the existence of such power, evidence has been offered of market share, barriers to entry, size, and the exercise of power''[53]. In 1987, Judge Greene attributed competitive problems to technology rather than to regulation:

> ''Local exchange competition has failed to develop, not so much because state and local regulators prohibit entry into the market by would-be competitors of the Regional Companies, but because of the economic and technological unfeasibility of alternative local distribution technologies''[54].

According to George J. Stigler (1968), barriers to entry are long-run costs imposed on entrants that are not present for incumbents[55]. Sunk costs are often said to be a barrier to entry if entrants need to make irreversible investments in capacity, while incumbents have already incurred these costs. The standard reasoning is as follows. The incumbent need only price to recover operating expenses and incremental capital expenditures, since the irreversible investment costs of entry can be written off. An entrant, in contrast, must anticipate earnings exceeding operating costs, incremental investment as well as the irreversible costs of establishing its facilities before it will decide to enter.

However, as Richard Posner (1976) points out, nonrecurring costs of entry are ''irrelevant if there are small firms in the market that can grow to be large firms.'' He continues ''In any event, there is grave doubt whether there are important nonrecurring costs of entry–barriers to entry in the true sense'' (Posner, 1976). He notes that economies of scale are not a barrier to entry since it is only a technological specification of efficient size. Costs of capital are not a barrier since the costs should be comparable for firms in the market. Finally, access to scarce inputs can be addressed as an input monopoly, which is the crucial issue in telecommunications.

Sunk costs are unlikely to be a barrier to entry particularly in industries involving substantial technological change such as telecommunications (Spulber, 1989; 1995). Any competitive market involves some degree of irreversible investment – whether in capital equipment, marketing, or research and development. Entrants commit capital resources in markets where they expect to earn competitive returns on their investments. The sunk costs involved in establishing a telecommunications system, given currently available technologies, are not qualitatively different from irreversible investments in any other competitive market. Entry into local exchange telecommunications particularly since the MFJ

[52] United States v. Western Elec. Co., 673 F. Supp. pp. 525, 538 (D.D.C. 1987).
[53] United States v. American Tel & Tel. Co., 524 F. Suppl. pp. 1336, 1346 (D.D.C. 1981).
[54] United States v. Western Elec. Co., 673 F. Supp. pp. 525, 537–38 (D.D.C. 1987).
[55] The definition is commonly applied, see for example Baumol and Willig (1981).

demonstrates that many companies are willing and able to make the irreversible investments required to enter into local telecommunications.

The notion that sunk costs are a barrier to entry depends in part on the similarity of the technology of the incumbent and entrant. The incumbent can write off its irreversible capital investment and price in a manner that covers the incumbent's operating cost. This would deter the entry of a competitor who must cover both operating cost and irreversible entry costs if the entrant's technology has a cost structure that is similar to that of the incumbent. The entrant need not be deterred from entry if it has a sufficient cost advantage over the incumbent. Technological change in telecommunications, such as the application of microprocessors in switching, has lowered the costs of operating networks. Moreover, by offering enhanced or specialised services, entrants can generate incremental revenues to offset the costs of entry. New technologies offer enhanced performance including switched services, the mobility of wireless services and the increased bandwidth of coaxial and fibre-optic systems.

Entrants can reduce the risk associated with making investment commitments in a variety of ways, including contracting with customers before irreversible investments are made, and entering into joint ventures or mergers with incumbents. Furthermore, in competitive markets there is often duplication of investment, and the entry of excess or insufficient capacity can take place as a consequence of uncertainty regarding costs, technology, or market demand. When entry creates excess capacity, this often leads to vigorous competition and industry shakeouts that reduce the profits of incumbents.

Evaluation of whether or not there are barriers to entry in telecommunications cannot be based on the technological characteristics and costs of the networks of the RBOCs because an entrant need not duplicate the RBOC's entire transmission and switching systems to enter the market profitably. The entrant need only enter portions of the market where the expected revenues exceed the costs of providing new service. Thus, a facilities-based entrant establishes a network to serve only its target customers. The FCC's analysis of entry patterns since the 1996 Act found that "there is a statistically significant and positive relationship between the probability a market is entered and the number of households in the area." They further show that "in all periods after 1996, the relationship between the percentage of the population in areas typically characterised by high business concentration, dense urban areas, and the probability the area is entered is statistically significant and positive"[56].

An entrant need not duplicate the technology of the incumbent RBOC's transmission and switching systems. The technology of telecommunications transmission has changed substantially in a manner that alters the types of investment required to establish a network. The transmission wires portion of

[56] Common Carrier Bureau, Local Competition: August 1999, Industry Analysis Division, Common Carrier Bureau, Federal Communications Commission, p. 8.

the traditional transmission technology represents the primary irreversible investment incurred by the local exchange carriers (LECs). In contrast, many new technologies (such as cellular, digital PCS, and satellite transmission) substantially reduce the wired portion of the transmission system. These new wireless technologies, by avoiding the wire-based transmission technology, substantially reduce transaction-specific irreversible investment. By serving areas using wireless transmission, particularly for the "last mile" to the customer's location, the new technologies are not tied to specific customers through investment in transmission.

The argument that sunk costs constitute a substantial barrier to entry into the local exchange is further rendered moot by the substantial entry into local telecommunications that began before the 1996 Act and has continued afterwards[57]. Although the market share of the competitive local exchange carriers (CLECs) and competitive access providers (CAPs) is small, they are almost ubiquitous. According to the FCC, "The geographic reach of facilities-based competition also continues to increase. By the end of June 1999, facilities-based CLECs were in every state, and in all but 18 of the nation's 193 local access and transport areas (LATAs)"[58]. By 1999, local competitors had at least 16% of fibre optic capacity for local calls[59]. In addition to CLECs, the reach of wireless providers is nearly universal and cable systems pass most households. The large number of small firms in the market and the presence of large firms with small market shares satisfies Posner's criterion for the irrelevance of entry barriers.

Whether or not the 1996 Act was instrumental in encouraging entry into local exchange telecommunications has yet to be determined. The process of interconnection may have been slowed by controversy over the FCC's interpretation of the access pricing provisions of the Act. The FCC applied the checklist provisions for the local exchange to restrict entry of RBOCs into long distance.

3.3. Tying

Under the tying doctrine, a company engages in leveraging monopoly power by requiring customers to purchase a bundle of goods. A company having market power in the sale of one good (the tying good) is said to leverage its market power

[57] Kellogg et al. (1999) observe that "Since divestiture, there has been a steady, inexorable rise in competition in all levels of the industry." They continue: "This process has accelerated rapidly since 1996 wherever competiiton makes strategic and economic sense for the new entrants. That competitive sphere includes business services of all kinds: short-haul toll services, wireless services, many data services, and other enhanced services. As of late 1998, nearly 300 companies were providing competitive local exchange carrier service of some description–companies like MCI WorldCom and AT&T/TCG, cable companies, interexchange carriers, providers of personal communications services (PCS), providers of shared tenant services (e.g., service to apartment buildings), and others. Well over 2000 interconnection agreements had been reached. New capital investments by competitors had surpassed new capital investment by the incumbents. Competitive local carriers had installed over 1000 switches, and were deploying new switches much faster than the incumbents."

[58] Common Carrier Bureau, Id. p. 5.

[59] Id. p. 2.

into the market for the second good (the tied good)[60]. A telecommunications company is alleged to engage in tying when it requires its customers to purchase bundles of services, some of which are monopolised because they are provided using that company's network and others of which are sold on competitive markets.

The concept of tying has been applied in antitrust allegations of leveraging. Because AT&T provided both local and long distance services as a bundle, this presumably leveraged its monopoly in local exchange telecommunications into the market for long distance. In a similar vein, providing telecommunication services and equipment manufacturing as a bundle presumably leveraged the monopoly in local exchange telecommunications into the market for equipment. This could apply to either customer premises equipment purchased by customers or network equipment that AT&T provided to itself.

Richard Posner observed that:

> "One striking deficiency of the traditional, 'leverage' theory of tie-ins, as the courts have applied it, is the failure to require any proof that a monopoly of the tied product is even a remotely plausible consequence of the tie-in"[61].

The problem is that the monopolist has little incentive to do so since it generally obtains no additional profits from the tie-in sale and rarely obtains market power in the market for the tied good. The firm may obtain benefits from tie-ins if the tied good allows metering of usage of the tying good for pricing purposes. However, as Posner observes, this does not imply that there is an incentive to exclude competition.

Given a monopoly in the tying good and a competitive market in the tied good, monopoly profits cannot be increased beyond those already available in selling the tying good. Creating a second monopoly in the tied good will not increase profits, since earnings are constrained by competitive alternatives in the market for the tied good for a firm selling complementary goods in fixed proportions. The local exchange company that provides access to the local exchange with each long distance call might be said to provide minutes of local network access in fixed proportion to minutes of long distance service.

[60] There is an extensive case law involving tying. The courts have applied Sherman Act Section 1 forbidding monopolisation to tying. The Clayton Act Section 3 forbids tying if it is anticompetitive and tends to create a monopoly. Alleged tie-ins include such things as complementary goods, services, lease agreements, and system components. In the 1998 case of US v. Microsoft, the DOJ accused Microsoft of tying its Explorer Internet Browser to its Windows operating system.

[61] See Posner (1976, p. 172). The reasoning behind the law and court decisions on tying has been subject to a significant amount of criticism. Bork (1978, p. 366) calls the transfer of market power theory behind the antitrust law on tying, and the Supreme Court's enforcement of it, "fallacious." See also Bowman (1957). For an overview of tying as a form of taxation see Katz (1989).

Michael Whinston (1990) shows that the monopolist does not have an incentive to leverage even with differentiated products in the tied-goods market, with economies of scale in production, and with oligopolistic, rather than a competitive, market structure in the tied-goods market. In his analysis, Whinston assumes that the tying good is an essential facility. He demonstrates that when there are two firms in the tied good market, it is not profitable for firm 1 to drive firm 2 out of business in the tied good market: "The key point is that with complementary products used in fixed proportions, the monopolist can actually derive *greater* profits when its rival is in the market than when it is not because it can benefit through sales of its monopolised product from the additional surplus that its rival's presence generates (due to product differentiation)"[62].

A company with a monopoly over local exchange services could benefit from increased demand for its services due to demand for complementary products offered by its competitors, such as long distance. It seems likely that the competitive provision of long distance, equipment and information services has led to increased demand for complementary local exchange services.

Under the Bell system, AT&T was able to effectively monopolise complementary products by the use of regulatory prohibitions against entry. For example, the company was able to prevent all attachments directly through regulation rather than requiring customers to purchase services as a bundle. Given the comprehensive nature of its network and the absence of competing networks, a market for complementary products was restricted by company tariffs supported by regulation. Thus, although AT&T sold bundles of services, monopolisation was due less to tying than to regulatory control of entry. AT&T's consistent opposition to entry into complementary products suggests that the company pursued a regulatory strategy aimed at maintaining its end-to-end regulated monopoly. AT&T likely viewed the entry of rivals into such areas as customer premises equipment as a competitive threat that would have begun to chip away at the Bell system monopoly. Such a regulatory strategy was pursued at the cost of company profits and consumer benefits that would have resulted from competitive entry into complementary products.

The process of unbundling of equipment and telecommunications services has its origins in two classic cases: Hush-A-Phone and Carterfone. Hush-A-Phone was an attachment that simply fitted over a telephone receiver and was designed to give some privacy to users whispering into the device[63]. AT&T argued that the device was unnecessary and that foreign attachments violated its tariffs. Although the FCC denied customers the right to attach the device, the D.C. Circuit Court reversed the FCC's decision in 1956.

[62] Note that the result depends on using the goods in fixed proportions. The result may not hold in the rather unlikely case in which there is a monopoly in the tying good and there is a single rival in the tied-goods market, and that rival produces an inferior good, see Whinston (1990).

[63] Hush-a-Phone Corp. v. U.S., 238 F.2d, p. 266 (D.C. Cir. 1956).

In 1959, Tom Carter invented a means of connecting mobile radio systems to the wireline network. AT&T declared that customers using Carterfone with an AT&T telephone would be subject to penalties under AT&T's FCC tariff number 132, which provided that: "No equipment, apparatus, circuit or device not furnished by the telephone company shall be attached to or connected with the facilities furnished by the telephone company, whether physically, by induction or otherwise." Carter filed an antitrust suit against AT&T and lost in the Fifth Circuit. The decision was reversed by the FCC, which found that AT&T's tariff was discriminatory because end users could install AT&T-manufactured equipment that did the same thing as Carterfone without the penalties. The decision on the right to interconnect became an FCC rule known as Part 68 and began the deregulation of customer premises equipment, affecting the development of a range of devices including mobile phones, modems, fax machines, and cable set-top boxes (Oxman, 1999). The Carterfone decision continues to guide FCC policy making on connection of equipment to telecommunications networks.

3.4. Cross subsidisation

Another form of alleged foreclosure is cross subsidisation. Such foreclosure is said to occur when the incumbent firm uses profits from a monopoly market to engage in *cross subsidisation* of other lines of business, thus gaining an unfair competitive advantage in other markets. Part of the problem in regulated industries is that the rate structure often contains cross subsidies at the behest of regulators. This practice can create economic inefficiencies since the customers of product A would be made better off if the products were produced and priced separately, even though this would forego economies from joint production. Moreover, the customers of product B are given incorrect price signals about incremental costs. Distorted prices also adversely affect incentives for competitors.

In MCI v. AT&T and Southern Pacific v. AT&T, plaintiffs alleged that AT&T engaged in cross subsidisation, using revenues from regulated services to lower prices on competitive services in a predatory fashion. In the MFJ proceedings, Judge Greene stated that provisions of the consent decree were designed to "(1) to prevent the divested Operating Companies from discriminating against AT&T's competitors, and (2) to avoid a recurrence of the type of discrimination and cross-subsidisation that were the basis of the AT&T lawsuit"[64]. Judge Greene stated that the RBOCs would be able "to subsidise the prices of their interexchange services with profits earned from their monopoly services" and "to subsidise the prices of their [information] services with revenues from the local exchange monopoly," and they "would have an incentive to subsidise the prices of

[64] United States v. AT&T, 552 F. Supp., pp. 131, 142 (D.D.C. 1982).

their equipment with the revenues from their monopoly services." These considerations led to the restrictions on RBOC entry into long distance under the 1996 Act.

Cross-subsidisation occurs when a company supplying more than one product or service uses the revenues from product A to recover a portion of the additional costs of producing product B. Consider a regulated firm that has a break-even rate structure[65]. A regulated firm's rate structure can be said to be free of cross-subsidies if and only if the prices satisfy the *stand-alone cost test*[66]. Stand-alone cost refers to the firm's long-run total cost of each service operated separately[67]. The stand-alone cost test requires that the revenues generated from either of two services not exceed the stand-alone cost of providing that service. If the revenues from one service do exceed its stand-alone cost, then that service is providing a cross-subsidy to the other service[68]. Clearly, the customers of the service that is providing the cross-subsidy would be better off if that service could be obtained independently of the other service.

A regulated firm's rate structure is free of cross-subsidies if and only if the prices satisfy the *incremental cost test*, which is equivalent to the stand-alone cost test for a break-even rate structure[69]. Applying the incremental cost test, revenues generated by each service must cover the incremental cost of providing that service[70]. The rationale for the incremental cost test is the requirement that each service must generate revenues that at least cover the additional cost of producing that service. If not, the other service is providing a cross-subsidy, and the customers of the other service would be better off receiving their service independently, at its stand-alone cost.

If the firm's rates are not necessarily break-even regulated rates but instead generate revenues that are greater than or equal to costs, then the incremental cost test should be applied to determine cross subsidisation. If the firm operates in both regulated and unregulated markets, the revenues in the unregulated market should cover the firm's incremental costs of serving the unregulated market. This

[65] For example, if the company produces two outputs, the company with a break-even rate structure has prices such that revenues equal costs, $p_1Q_1 + p_2Q_2 = C(Q_1, Q_2)$.

[66] See Baumol, Panzar and Willig (1982, pp. 352–353).

[67] For example, for a multiproduct cost function involving two outputs, the stand-alone costs of outputs 1 and 2 are given by $C(Q_1, 0)$ and $C(0, Q_2)$ respectively.

[68] The stand-alone cost test with two products is as follows. Revenues must not exceed stand alone cost for either of the products, $p_1Q_1 \leqslant C(Q_1, 0)$ and $p_2Q_2 \leqslant C(0, Q_2)$. In the case of more than two products, the test requires that no group of products subsidizes any other group of products.

[69] Incremental cost in the case of two products equals $C(Q_1, Q_2) - C(0, Q_2)$ for product 1 and $C(Q_1, Q_2) - C(Q_1, 0)$ for product 2. For further discussion and formal definitions, see Baumol (1986, pp. 113–120) and Spulber (1989, Chapter 3).

[70] The incremental cost test for only two products is defined by $p_1Q_1 \geqslant C(Q_1, Q_2) - C(0, Q_2)$ and $p_2Q_2 \geqslant C(Q_1, Q_2) - C(Q_1, 0)$. In the case of more than two products, the revenues generated by each group of products must cover the incremental cost of providing that group of products.

guarantees that serving the unregulated market increases or does not reduce the firm's profit.

Judge Greene specified however that demonstrating foreclosure would require showing that the RBOCs have both "the incentive and opportunity to act anticompetitively"[71]. Cross subsidisation cannot long survive in a competitive market because companies will have an incentive to enter into those markets in which incumbents are generating subsidies. As Judge Greene observed in 1982: "AT&T's opportunity for cross subsidisation will become increasingly curtailed as interexchange competition increases"[72]. As regulated markets are opened to competition, companies typically cannot sustain regulated rate structures with cross subsidies since overpriced goods are subject to intense competition and underpriced goods generate losses. Moreover, profit-maximising firms have little incentive to engage in cross-subsidisation since they do not wish to maintain lines of business that continue to generate revenues that are less than incremental costs.

4. Mergers

In the competitive environment since the 1996 Telecommunications Act, competition policy in telecommunications is perhaps best defined not by what it does but by what it does not do. The choice of antitrust policy makers not to challenge the series of large-scale telecommunications mergers since the 1996 Act indicates not so much forbearance as a recognition of the strength of competition that has developed in this sector. Although economists have considered collusion issues in telecommunications, competition policy also has not shown much concern with collusion in telecommunications perhaps because of the increasingly competitive nature of telecommunications markets[73].

In addition to the general merger enforcement activities of the DOJ and the Federal Trade Commission (FTC), the FCC reviews telecommunications mergers. The FCC may object to mergers or place conditions for approval of the mergers. The FCC's merger activities have generated some controversy. For example, Senator John McCain, Chairman of the Committee on Commerce, Science,

[71] United States v. American Tel. & Tel. Co., 552 F. Supp., pp. 131, 187–190 (D.D.C. 1982).

[72] United States v. American Tel. & Tel. Co., 552 F. Supp., p. 131, (D.D.C. 1982) here p. 187

[73] MacAvoy (1995) examines price-cost margins and market shares in long distance during the decade after the 1984 implementation of the MFJ. He shows that the price-cost margins for AT&T, MCI and Sprint increased for seven important classes of long-distance services and argues that such price movements were consistent with tacit collusion. The continued exclusion of RBOCs from long distance markets suggests that regulatory concerns about the local exchange outweigh any potential concerns about additional entry in long distance markets. Laffont, Rey, and Tirole (1998) and others suggest that interconnection prices could potentially function as a collusive device if two or more companies operating interconnecting networks agree on access charges as a means of raising final prices.

and Transportation, observed that in contrast to the DOJ or the FTC, the "third agency, the Federal Communications Commission, has comparatively little expertise in these issues, and only limited authority under the law." McCain continued:

> "Nevertheless, the FCC has bootstrapped itself into the unintended role of official federal dealbreaker. How? by using its authority to impose conditions on the FCC licenses that are being transferred as part and parcel of the overall merger deal. Because the FCC must pre-approve all license transfers, its ability to pass on the underlying licenses gives it a chokehold on the parties to the merger. And it uses that chokehold to prolong the process and extract concessions from the merging parties that oftentimes have very little, if anything, to do with the merger itself."

Senators John McCain, Orrin Hatch, Chairman of the Judiciary Committee, and John Ashcroft introduced a bill (S.1125) that joined a spate of bills seeking to limit the FCC's powers of merger review[74].

Antitrust challenges to mergers have been limited. For example, in U.S. v. SBC and Ameritech, the DOJ sought relief under Section 7 of the Clayton Act and the District Court for the District of Columbia required the companies to divest overlapping cellular units. In U.S. v. AT&T and TCI, the DOJ sought relief under Section 7 of the Clayton Act. The District court for the District of Columbia ordered TCI to divest itself of Liberty Media's Sprint holdings as a condition of the AT&T/TCI merger[75].

A spate of *horizontal mergers* in long-distance telecommunications illustrates antitrust policy toward that market. WorldCom grew large through a series of over 70 acquisitions. WorldCom bought the largest competitive access provider, MFS Communications, for approximately $12 billion (Landler, 1996). Thus, WorldCom will become a fully vertically integrated local exchange and long distance carrier. The merger demonstrates the importance of "one-stop shopping," with one company providing a package of local and long distance services, as well as data transmission and Internet connections. Moreover, the company was able to provide end-to-end transmission using both local transmission facilities and long distance transmission facilities. WorldCom also acquired the second largest long distance company MCI Communications for approximately $42 billion to form MCI WorldCom, the second-largest long-distance carrier. The consolidation of the long distance industry through the WorldCom-MCI merger created a large rival for AT&T. The absence of challenges to these mergers recalls the FCC's conclusion that AT&T was not a dominant carrier.

The FCC approved WorldCom's MCI acquisition subject to two conditions: MCI would divest its Internet assets and the transfer of MCI's direct broadcast satellite license to WorldCom would be subject to the already existing

[74] Senator John McCain, Press release, May 26, 1999, Washington, D.C.
[75] See http://www.usdoj.gov/atr/cases.html.

review process. The Commission found the merger would be "consistent with the 'pro-competitive, de-regulatory' framework of the Telecommunications Act of 1996 and may produce tangible benefits to consumers." The FCC argued that the combination of WorldCom's extensive local exchange facilities, small and medium business customer base, and foreign networks and MCI's national brand name, marketing expertise, and broad residential base would hasten entry into local telephone markets. The FCC concluded that although the merger combined the second and fourth largest long distance companies, "the trend toward increased long distance competition is expected to continue in light of the construction of new long distance networks." The FCC further felt that "the merger will create a stronger local service competitor than either of the companies would provide absent the merger"[76].

MCI WorldCom attempted to buy the number three carrier Sprint for approximately $115 billion but was forced by regulators to abandon the merger. Attorney General Janet Reno stated that "This merger threatens to undermine the competitive gains achieved since the department challenged AT&T's monopoly 25 years ago"[77]. It was viewed as likely that WorldCom will attempt to acquire additional long distance companies such as Qwest Communications International, while both WorldCom and Sprint were seen as potential takeover targets for international telecommunications companies such as Deutsche Telekom, France Télécom, British Telecommunications and Nippon Telegraph and Telephone[78].

Mergers among local exchange carriers have reduced the number of RBOCs from seven to four. These mergers were not viewed as horizontal since the companies serve different territories. Bell Atlantic acquired Nynex for about $19.5 billion and later acquired GTE (not one of the RBOCs) for over $52 billion to form Verizon Communications. The merger between Bell Atlantic and Nynex was approved subject to conditions imposed by the FCC governing interconnection with other carriers and performance in serving retail customers and other carriers[79]. Similarly, the FCC approved the merger between Bell Atlantic and GTE subject to conditions regarding long distance and Internet service. SBC Communications purchased Pacific Telesis Group for over $16 billion, Ameritech for $75 billion, and Southern New England Telecommunications Corp. for $4.4 billion. The FCC approved the Ameritech merger subject to 30 conditions designed to open their local markets to further competition, stimulate the deployment of advanced broadband services, and to strengthen the merged firm's incentives to expand competition outside its 13 state service area. In addition, the

[76] FCC Approves Merger of WorldCom and MCI (CC Docket No. 97–211), FCC Press release, Washington, D.C., September 14, 1998.
[77] Quoted in Schiesel and Leonhardt, 2000, p. A1.
[78] Romero, 2000, p. C5.
[79] See Memorandum Opinion and Order, 12 FCC Rcd 19985, App. C (1997) ("Merger Order").

Internet network operator Quest Communications International acquired U. S. West for over $37 billion[80].

A series of acquisitions in the cable industry were conglomerate mergers with vertical elements[81]. AT&T acquired the cable companies Tele-communications Inc and Media One Group for $52 billion and $55 billion respectively. The AT&T acquisition combined cable systems across service areas where the companies generally did not compete so the merger did not involve horizontal consolidation. The leading Internet service provider America Online (AOL) acquired the media conglomerate Time Warner, which is the second largest cable company. The FTC and the FCC approved the AOL-Time Warner merger subject to conditions on Internet access.

In United States v. Primestar, the Division challenged and subsequently blocked Primestar's acquisition of the direct broadcast satellite ("DBS") assets of News Corporation Limited and MCI alleging that since five of the largest cable companies control Primestar, the acquisition would lessen competition in video distribution, including cable and DBS. Primestar would have acquired from News Corp./ MCI the operation of 28 satellite transponders at the 110 west longitude orbital slot and two high-power DBS satellites[82].

5. Conclusion

Antitrust policy seeks to set rules for competitive behaviour. This function is limited in heavily regulated markets. Accordingly, public and private antitrust enforcement has tended to follow regulatory policy in telecommunications. Significant technological change, diverse transmission technologies, new types of services, and substantial entry define the new era in telecommunications since the break up of the Bell system. As regulation has receded, the emphasis of antitrust in telecommunications has changed substantially.

[80] FCC Approves SBC-Ameritech Merger Subject to Competition Enhancing Conditions, FCC Press Release, Washington, D.C., October 6, 1999.

[81] In United States v. US West, Inc. and Continental Cablevision, Inc., the Antitrust Division challenged the $11.8 billion merger between US West and Continental Cablevision, the third largest cable system operator in the United States, and simultaneously filed a proposed consent decree requiring Continental Cablevision to divest its interest in Teleport Communications Group, see Twentieth Annual Report Pursuant to Subsection (j) of Section 7A of the Clayton Act Hart-Scott-Rodino Antitrust Improvements Act of 1976 Department of Justice and Federal Trade Commission, Fiscal Year 1997.

[82] United States v. Primestar, Inc., Tele-Communications, Inc., TCI Satellite Entertainment, Inc., Time Warner Entertainment Company, L.P., MediaOne Group, Comcast Corporation, Cox Communications, Inc., GE American Communications, Inc., Newhouse Broadcasting Corporation, The News Corporation Limited, MCI Communications Corporation and Keith Rupert Murdoch. See Twenty First Annual Report, Pursuant to Subsection (j) of Section 7A of the Clayton Act Hart-Scott-Rodino Antitrust Improvements Act of 1976 Department of Justice and Federal Trade Commission, Fiscal Year 1998.

The first two major government antitrust actions against AT&T focused on changing conduct through negotiated settlement. The third government suit ended in a structural remedy, breaking up the firm, but featured conduct restrictions on the RBOCs that perpetuated regulation rather than relying on markets. The Telecommunications Act of 1996 took antitrust policy out of the business of regulation, although the Department of Justice continues to advise the FCC on RBOC entry into long distance and to condition telecommunications mergers.

The industry has changed from a simple duet of telegraphy and telephony to a symphony of interconnected alternatives including wireline telephony, wireless telephony, cable television and telephony, satellite communications, private data networks and the Internet. Telecommunications has gone from an industry dominated by a single firm to a sector of the economy with many giant firms and thousands of smaller companies. What was once a vertically integrated industry has moved toward significant vertical separation, with separate industry groups providing many different types of transmission facilities and services, a wide variety of specialised equipment manufacturing, and a proliferation of information services. With such diverse markets and many different firms, antitrust policy can return to its intended role as an impartial arbiter of competition.

References

Areeda, P. and D. F. Turner, 1975, Predatory pricing and related practices under Section 2 of the Sherman Act, Harvard Law Review, 88, 697–733.

Baumol, W. J., 1986, Superfairness: Applications and theory, Cambridge, MA: MIT Press.

Baumol, W. J. and J. G. Sidak, 1994, Toward competition in local telephony, Cambridge, MA: MIT Press.

Baumol, J. C. Panzar and R. D. Willig, 1982, Contestable markets and the theory of industry structure, New York: Harcourt Brace Jovanovich; rev. ed. 1988.

Baumol, W. J. and R. D. Willig, 1981, Fixed cost, sunk cost, entry barriers and sustainability of monopoly, Quarterly Journal of Economics, 95, 405–431.

Bork, R. H., 1978, The antitrust paradox: a policy at war with itself, New York: Basic Books.

Bowman, W. S., 1957, Tying arrangements and the leverage problem, Yale Law Journal, 67, 19–36.

Breyer, S. J., 1982, Regulation and its reform, Cambridge, MA: Harvard University Press.

Calhoun, G., 1992, Wireless access and the local telephone network, Boston, MA: Artech House.

Crandall, R. W., 1991, After the break-up: U.S. telecommunications in a more competitive era, Washington, D.C.: Brookings Institution.

Demsetz, H., 1968, Why regulate utilities, Journal of Law and Economics, 11, 55–65.

Evans, D. S. and J. J. Heckman, 1984, A test for subadditivity of the cost function with an application to the Bell system, American Economic Review, 74, 615–623.

Grosvernor, E. S. and W. Morgan, 1997, Alexander Graham Bell: The life and times of the man who invented the telephone, New York: Harry N. Abrams.

Horwitz, Robert, B., 1989, The irony of regulatory reform: the deregulation of American telecommunications, Oxford: Oxford University Press.

Huber, P. W., 1997, Law and disorder in cyberspace: Abolish the FCC and let common law rule the telecosm, Oxford: Oxford University Press.
Katz, M. L., 1989, Vertical contractual relations, in R. Schmalensee and R. D. Willig, eds., Handbook of Industrial Organisation, Vol. 1, Amsterdam: North-Holland.
Kellogg, M. K., J. Thorne and P. W. Huber, 1992, Federal telecommunications law, Boston: Little, Brown.
Kellogg, M. K., J. Thorne and P. W. Huber, 1999, Federal telecommunications law, Second edition, Aspen: Aspen Publishers.
Krattenmaker, T. G. and S. C. Salop, 1986, Anticompetitive exclusion: Raising rivals' costs to achieve power over price, Yale Law Journal, 96, 297–326.
Laffont, J-J., P. Rey, and J. Tirole, Network Competition: I. Overview and nondiscriminatory pricing, RAND Journal of Economics, 29, 1–37.
Landler, M., 1996, WorldCom to buy MFS for $12 billion, Creating a phone giant, New York Times, C1.
Lopatka, J. E. and P. E. Godek, 1992, Another look at Alcoa: Raising rivals' costs does not improve the view, Journal of Law and Economics, 35, 311–330.
Lott, J. R., 1999, Are predatory commitments credible? Who should the courts believe? Chicago: University of Chicago Press.
MacAvoy, P. W., 1995, Tacit collusion under regulation in the pricing of interstate long distance services, Journal of Economics and Management Strategy, 4, 147–185.
McGee, J. S., 1958, Predatory price cutting: The Standard Oil (N.J.) case, Journal of Law and Economics, October, 137–169.
McGee, J. S., 1980, Predatory pricing revisited, Journal of Law and Economics, 23, 289–330.
Mill, J. S., 1848, Principles of political economy, Vol. 1, New York: A. M. Kelly, reprinted 1961.
Noll, R. G. and B. M. Owen, 1989, The anticompetitive uses of regulation: United States v. AT&T, in J. Kwoka and L. White, eds., The Antitrust Revolution, Oxford: Oxford University Press.
North American Telecommunications Association, 1991, Industry basics, Fourth edition.
Ordover, J. A. and G. Saloner, 1989, Predation, monopolisation and antitrust, in R. Schmalensee and R. D. Willig, eds., Handbook of Industrial Organisation, vol.1, 537–596.
Oxman, J., 1999, The FCC and the unregulation of the Internet, Office of Plans and Policy, OPP working paper no. 31, Washington D.C.: Federal Communications Commission.
Phillips, Jr., C. F., 1988, The regulation of public utilities: theory and practice, Arlington, VA: Public Utilities Reports, Second edition.
Posner, R. A., 1976, Antitrust law: an economic perspective, Chicago: University of Chicago Press.
Reiffen, D. and A. N. Kleit, 1990, Terminal railroad revisited: foreclosure of an essential facility or simple horizontal monopoly, Journal of Law and Economics, 33, 419–438.
Romero, S., 2000, WorldCom and Sprint end their $115 Billion Merger, New York Times, July 14. p. C5.
Salop, S. C. and D. T. Scheffman, 1983, Raising rivals' costs,American Economic Review, Papers and Proceedings, 73, 267–271.
Schiesel, S., 2000, Justice Dept. Opposes SBC in Texas Bid, New York Times, February 15, 2000, p. C8.
Schiesel, S. and D. Leonhardt, 2000, Justice Dept. acts to block merger of phone giants, New York Times, June 28, A1.
Shin, R. T. and J. S. Ying, 1992, Unnatural monopolies in local telephone, RAND Journal of Economics, 23, 171–183.
Sidak, J. G. and D. F. Spulber, 1998, Cyberjam: Internet congestion of the telephone network, Harvard Journal on Law and Public Policy, 21, 327–394.
Spulber D. F., 1995, Deregulating telecommunications, Yale Journal on Regulation, 12, 1995, 25–67.
Spulber, D. F., 1989, Regulation and markets, Cambridge, MA: MIT Press.
Stigler, G. J., 1968, Barriers to entry, economies of scale and firm size, in The Organisation of Industry, Homewood, IL: Irwin.

Temin, P. with L. Galambos, 1987, The fall of the Bell system, New York: Cambridge University Press.
Vogelsang, I., 1997, Antitrust vs. sector specific approaches in regulating telecommunications markets, in D. Elixmann and P. Kürble, eds., Multimedia - Potentials and Challenges from an Economic Perspective, papers presented at the WIK Workshop, Königswinter, 3–4 December 1996, WIK Proceedings, No. 5, June.
Waldman, D. E., 1986, The economics of antitrust: Cases and analysis, Boston: Little, Brown.
Walras, L., 1936, Etudes d'economie politique appliquee: Theories de la production de la richesse socialle, Lausanne: F. Rouge.
Whinston, M., 1990, Tying, foreclosure and exclusion, American Economic Review, 80, 837–859.

Chapter 12

COMPETITION IN THE LONG-DISTANCE MARKET

DAVID L. KASERMAN

Auburn University

JOHN W. MAYO

Georgetown University

Contents

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

Few, if any, topics within the telecommunications policy arena have been more hotly debated than the question of competition in the long-distance market. Indeed, the separation of AT&T from the Bell Operating companies in 1984, which, at the time, was the largest corporate divestiture in history, was prompted by ongoing debates regarding the potential for and impediments to long- distance competition. Rather than ending these debates, however, the divestiture created a new wave of issues surrounding the emergence and intensity of competition in this important segment of the telecommunications industry. Was competition in the long-distance market feasible or did the market retain natural monopoly characteristics that would ultimately doom hopes for emerging competition? When might competition be judged sufficient to relax the traditional regulatory constraints imposed on AT&T? Would the emergence of competition undermine the widely acknowledged goal of promoting universal telephone service? Would competition emerge selectively in urban areas and for large business customers only, leaving rural and residential consumers captive to the incumbent monopoly supplier? Would competition be able to develop despite the 'dominant' position of AT&T in the market?

These and numerous other questions have occupied the attention of a growing number of scholars over the past decade and a half. The impetus for research in this area has stemmed from two sources. The first is a natural academic interest in understanding the causes and consequences of one of the largest corporate and regulatory restructurings in history. The second is a corresponding surge in the demand for informed expert witnesses to appear in scores of regulatory hearings at both the state and federal levels as the governmental agencies involved in this process have struggled with the myriad policy questions raised by the restructuring.

The resulting research has spawned a burgeoning literature that has created two distinct benefits. First, this body of research has provided a considerably more sophisticated understanding of the nature of competition in the long-distance marketplace and a correspondingly improved basis upon which policy makers can fashion more effective telecommunications policies. Second, the scrutiny of long-distance telecommunications competition has proven to be a fertile field in which to test (and thereby improve) our general understanding of some of the fundamental theories of industrial organization.

In this chapter, we explore this literature with an eye toward identifying the areas in which particular advances have occurred and areas that remain in need of further analysis. The lessons to be gleaned from these prior studies have ramifications for a number of important policy issues. Most prominently, the intensity and durability of long-distance competition has been a central issue in debates regarding both deregulatory policies toward AT&T, and, more recently, the benefits of allowing the Regional Bell Operating Companies (RBOCs) to enter

their 'in-region' interLATA markets under the provisions of Section 271 of the Telecommunications Act of 1996[1]. Specifically, the greater the degree of competition in long distance, the stronger is the case for deregulation of that market and the weaker is the case for allowing RBOC entry, *ceteris paribus*. As a result, debate on this issue has been intense and, at times, contentious.

In addition to these two direct policy issues, the question of long-distance competition has other, less immediate (but equally far-reaching), applications. For example, to the extent competition has grown in the long-distance market, can we expect it to infiltrate successfully the more recently opened local exchange markets? That is, to what extent are our experiences in long distance likely to carry over to local exchange markets, and what suggestions do these experiences offer for public policy? Also, what, if any, lessons are there for emerging competition in other network industries (e.g. electricity and natural gas?) In particular, should the circuitous path that moves from vertically integrated regulated monopoly through the divestiture route to, hopefully, vertically integrated unregulated competition be followed in these industries as well? Similarly, within telecommunications, should other countries that are just starting down that path pursue this roundabout approach or should a more direct approach that does not require vertical divestiture as a starting point be used? Obviously, our review cannot answer all of these questions definitively; but we hope it will help to shed light on some of the important issues that such answers ultimately must resolve.

For organizational purposes, it is useful to provide a taxonomy of the individual studies that have arisen according to the basic methodological approaches they have employed to assess the degree of competition in long distance. Specifically, four fundamentally different approaches have been adopted. First, several authors have applied the traditional structure – conduct – performance approach of industrial organization economics. Second, empirical analyses of the observed impact of relaxed regulation of AT&T on market prices have been used by several authors to infer the presence or absence of competition. Third, several authors have attempted to draw inferences regarding the intensity of competition from observed relationships between output price changes and changes in carrier access charges, which constitute an important input price. Fourth, the more modern empirical methodology of residual demand estimation and oligopolistic interdependence (the so-called New Empirical Industrial Organization or NEIO studies) has been employed in several papers. In the sections that follow, we survey the results pertaining to each of these four methodologies. Finally, while the majority of work has focused upon the competitiveness of domestic long distance, interexchange calls placed within the U.S., there have been several recent studies of the competitiveness of international calls, calls originating in the U.S. and

[1] For a recent discussion of this debate, see Schwartz (2000). See, also, Beard, Kaserman, and Mayo (1999).

terminating in other countries. It is generally conceded that this segment of the long-distance market is likely to exhibit relatively less competitive outcomes due, *inter alia*, to comparatively greater incumbency advantages enjoyed by AT&T at divestiture. As a result, these studies may be viewed as providing a lower bound estimate of the intensity of competition in the industry generally. Accordingly, our final substantive section reviews this set of studies. A concluding section summarizes our findings.

2. Structure – conduct – performance

At least since the work of Bain (1956), industrial organization analyses have drawn upon the structure – conduct – performance (S-C-P) paradigm[2]. Broadly, this paradigm suggests that the structure of an industry will lead to predictable firm conduct within the industry, and that this conduct will, in turn, generate predictable economic performance. Within the construct of the S-C-P paradigm, then, it is argued that a detailed analysis of an industry's structure, broadly defined, can provide key insights into the extent of competition likely to prevail in the relevant market.

2.1. Structure

2.1.1. Market definition

The first step in any analysis of competition in an industry is to properly define the product and geographic dimensions of the relevant market. If a market is defined either too broadly or too narrowly, spurious conclusions may arise. In the case of voice telecommunications, the standard economic approach to market definition suggests that the relevant market for the purpose at hand consists of interexchange (long-distance) services sold in the United States[3]. The principal reason for the adoption of this broad product market definition is that the degree of, especially, supply-side substitution among providers of the various alternative long-distance services is quite high[4]. That is, firms that provide standard Message

[2] For a thorough generic assessment of the strengths and weaknesses of this approach, see Schmalensee (1989).

[3] Long-distance services in the U.S. have traditionally been distinguished from local exchange services by the presence of a "per-minute" charge. Since the divestiture, such long-distance calls have been divided into interLATA or intraLATA calling. For a more detailed description of these distinctions, see Kaserman and Mayo (1995, Chapter 18).

[4] There is also some (albeit limited) degree of demand-side product substitutability across individual services as well. See Taylor (1994) and his chapter on "Customer Demand Analysis" in this *Handbook*. Very little, if any, demand-side geographic substitution exists in the use of long-distance services. That is, a call from Washington, D.C. to Boston is not substitutable for a call from Washington, D.C. to San Francisco.

Toll Service (MTS) can very easily utilize the same telecommunications network to provide alternative long-distance services (e.g. WATS or 800 services). Indeed, many of these separate services are simply different pricing mechanisms applied to otherwise identical electronic signals being transmitted between two (or more) points. In its consideration of policy matters, the Federal Communications Commission (FCC) has adopted this single-product market definition, stating that it includes "all interstate, domestic, interexchange services... with no relevant submarkets"[5].

2.1.2. Concentration in the U.S. long-distance market

Under the S-C-P methodology, once the relevant market is defined, the structure of the market in question must then be scrutinized. Several dimensions of market structure are generally considered to be relevant. The number of competitors and the distribution of their market shares are potentially quite important. Data on these market concentration characteristics are readily available for the interstate segment of the long- distance market from the FCC. For example, in Figure 1, we track the number of long-distance providers over time[6]. This series shows that the number of long-distance carriers has grown dramatically in the period following the divestiture of AT&T, almost tripling over the 1987–1997 period[7].

The growth of new carriers in the long-distance market is undoubtedly due to several factors. Perhaps most notable among these was the consent decree (known as the Modification of Final Judgment, or MFJ) that led to the divestiture of the Bell Operating Companies from AT&T. Together with several policy actions taken by the FCC, the MFJ led to dramatically reduced legal and regulatory barriers to entry into the long-distance market[8]. For instance, the MFJ required that the Bell Operating Companies provide access to their facilities to all inter-exchange and information services providers that "is equal in type, quality, and price" to that provided to AT&T. This 'equal access' to all interexchange carriers was promoted by the FCC through requirements that local exchange carriers upgrade their facilities to permit customers to access their long-distance carrier of choice by simply dialing 1+. Customers, in turn, were asked to 'pre-subscribe' to

[5] See Order, In the Matter of Motion of AT&T Corp. to be Reclassified as a Dominant Carrier," Federal Communications Commission, released October 23, 1995, para. 22.

[6] Source: Table 10.2 in FCC "Trends in Telephone Service," September 1999.

[7] These competitors include both firms that provide long-distance service primarily over their own facilities (so-called facilities-based carriers) and firms that lease the underlying transmission facilities of other firms and combine that capacity with their own inputs to provide retail-stage services (so-called resellers). While it is sometimes asserted that resellers provide less competitive discipline on the market than facilities-based firms, it has also been argued that resellers are the source of considerable pricing discipline in the market. See Kaserman and Mayo (1990).

[8] For a description of the divestiture and the events that preceded the consent decree, see *inter alia*, Temin (1987); Faulhaber (1987); Brock (1981); Kaserman and Mayo (1995), and the "Historical Overview" by Brock in this *Handbook*.

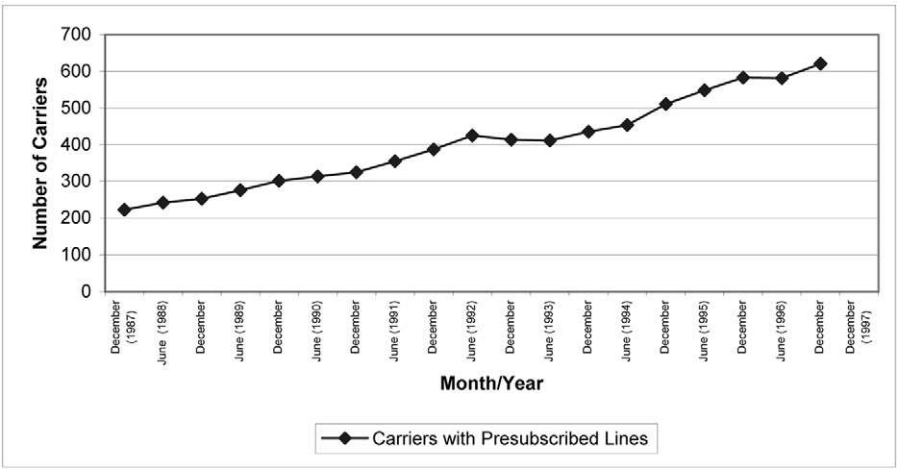


Fig. 1. Firms competing in the long-distance market.

their preferred long-distance carrier. The implementation of the equal access provisions led to the elimination of a dialing disparity for customers of AT&T's competitors – the so-called other common carriers (OCCs). Specifically, with equal access, customers of companies such as Sprint or MCI could dial 1+ rather than a series of additional digits to have the call carried by these alternative carriers. The upshot was a more homogeneous set of product offerings and, over time, a general trend toward convergence of the prices charged.

Another likely reason for the pronounced growth in the number of long-distance carriers has been the extraordinary growth rate in the demand for long-distance services. While real GDP grew by an average rate of 3.3 percent per year over the 1985–2000 period, the total number of interstate access minutes sold grew by 7.8 percent. The total output of long-distance services (measured in access minutes) is shown in Figure 2[9].

Prior empirical studies have found that, in general, demand growth is a powerful attractor to new market entrants[10]. With industry growth, the prospect for niche entrants to enter and succeed in the market expands. Many of the firms that have entered the long-distance market have focused their efforts on specific regional areas or on specific types of customers.

While both the reduction in regulatory barriers to entry and the rapid growth of demand for long-distance services has facilitated entry into the

[9] Source: Table 11.1 of FCC "Trends in Telephone Service," September 1999. The observed growth in the total size of the long-distance market is a function of several factors, including both price declines and a variety of non-price factors.

[10] See Hause and DuRietz (1983) and Burton, Kaserman and Mayo (1999).

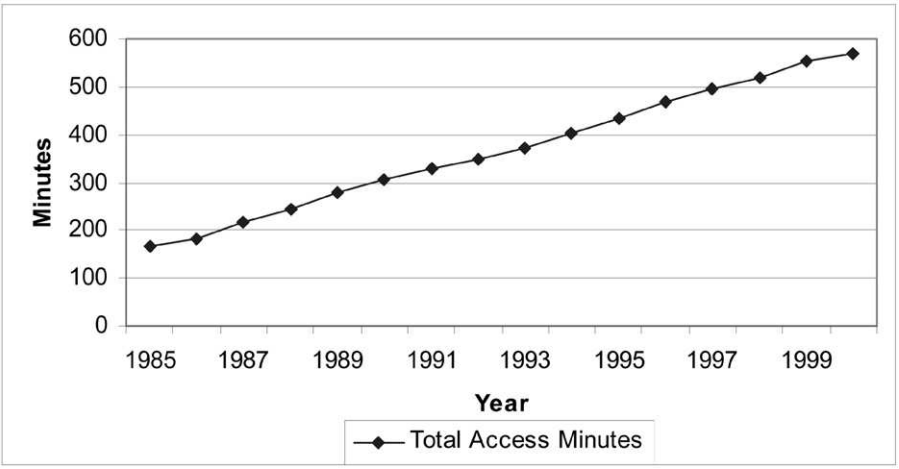


Fig. 2. Total Industry access minutes.

market, Taylor and Zona (1997, p. 232) suggest that yet another source of the observed entry by new firms may be at work in this market. In particular, they argue that

> "in a competitive market we should see little net entry, that is, the number of firms that exit (or the capacity of firms that exit) should be replaced by an equivalent number of firms or capacity in the long run."

Drawing on entry data such as those reported in Figure 1 above, Taylor and Zona then conclude that the persistence of observed entry into the long-distance market is evidence of above normal economic profits brought about by noncompetitive behavior. Given the likely role that relaxed regulatory barriers to entry and market expansion have had on the rapidly growing number of firms, however, the ability to infer noncompetitive performance from the observation of positive net entry into the long-distance market appears somewhat strained. While economic theory certainly indicates an attraction of prospective entrants into markets where incumbents behave non-competitively, the prospect that other *competitive* market forces are causing observed entry cannot be ignored.

In a similar vein, a host of pricing distortions inherited from over 50 years of regulation, as well as the continued regulation of this market, appears likely to have presented potential entrants with pockets of profitable entry opportunities. In particular, the presence of pervasive cross-subsidies has undoubtedly created incentives for entry into those market segments where the subsidies were being collected. Thus, entry incentives may be created by either a failure

of competitive market forces or distortions created by regulation (or, of course, both).

The divestiture, together with adoption of more liberal entry policies by federal and state regulators, gave rise not only to a host of new competitors, but also a corresponding erosion in the market share of the former regulated monopolist, AT&T. Figure 3[11] reveals the evolution of AT&T's market share (measured alternatively in minutes and revenues) over time[12].

In this figure, we see that, after beginning the post-divestiture period with roughly 90 percent of the interexchange marketplace, AT&T's market share has steadily fallen. By the end of 2000, AT&T controlled less than 40 percent of U. S. long-distance revenues. This marked erosion of AT&T's market share has come as a consequence of market share gains not only by its two largest competitors, MCI WorldCom and Sprint, but also by other, so called third-tier competitors.

A commonly used measure of overall market concentration is the Herfindahl-Hirschman Index (HHI). For the long-distance industry, an examination of the HHI reveals several interesting observations. First, as seen in Figure 4, the industry, which began with AT&T in control of roughly 90 percent of the market, generated extremely high concentration numbers in the early post-divestiture period[13].

Following divestiture, the HHI remained over 5000 until 1989. Second, there was a reasonably steady erosion in the degree of concentration of the inter-exchange marketplace from 1984 through 1997[14]. By 1997, the HHI for all firms providing long-distance services had fallen to just over 2000. Third, the 1998 data reveal the first increase in concentration in the long-distance market since the divestiture[15]. It seems reasonable to conclude that, while the long-distance industry remains 'highly concentrated' according to the Department of Justice and Federal Trade Commission's *Horizontal Merger Guidelines* (1992), the level of industry concentration has fallen markedly.

2.1.3. Barriers to entry

Another critical dimension of industry structure centers on entry conditions into the market. Indeed, the central role that entry conditions play in determining

[11] Source: Tables 11.1 and 11.3 of FCC "Trends in Telephone Service." September, 1999.

[12] In some industries, arguably including the long-distance industry, capacity may provide the most useful metric along which to measure market share. For a discussion, see Hovenkamp (1990). Nevertheless, we adopt the more frequently used market share measures in the long-distance industry here.

[13] Source: Tables 11.3 and 11.4 of FCC "Trends in Telephone Service," September, 1999.

[14] The erosion of the HHI reported here is in contrast to the stabilization of the HHI that is reported in MacAvoy (1995). The reason for the disparity is that MacAvoy calculates the HHI based only upon the largest three carriers, ignoring all other market participants.

[15] The increase in concentration in 1998 was in considerable measure due to the merger of MCI and WorldCom.

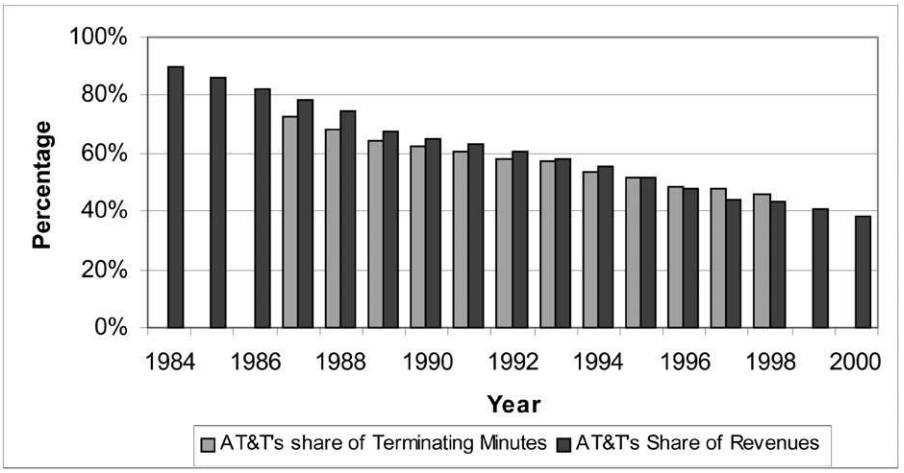


Fig. 3. AT&T's market share.

market performance is emphasized in the extant industrial organization literature. Scherer and Ross, for example, note that "[S]ignificant entry barriers are the sine qua non of monopoly and oligopoly" (1990, p. 11).

Two general avenues exist by which to determine the extent of barriers to entry into a market. First, it is possible to examine the 'laundry list' of potential barriers to entry that have been identified by economic theory as potentially operating to

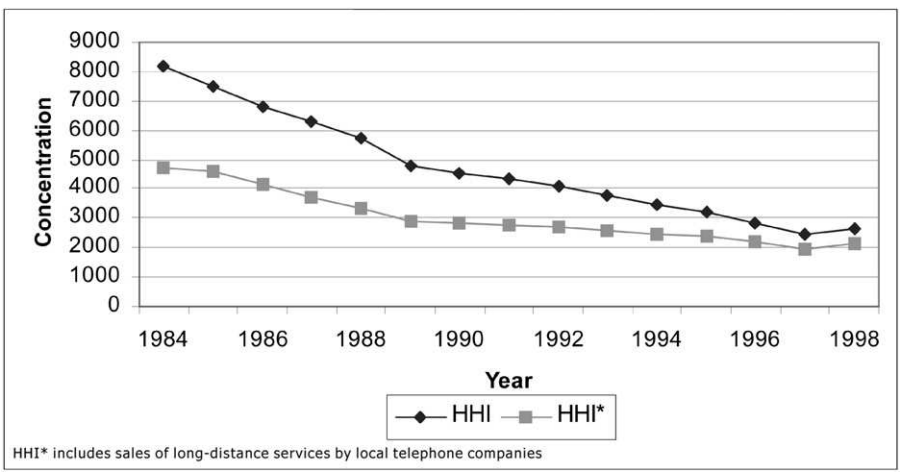


Fig. 4. Herfindahl Hirschman index.

impede entry into a market. At this point, the laundry list is quite long but would include, at a minimum, the prospect that entry may be deterred as a consequence of high capital costs, a high degree of, *inter-alia*, product differentiation, absolute cost advantages, large sunk costs, and switching costs. A second more direct approach is to examine the actual entry and exit patterns observed in a market to see what inference might be drawn regarding the height of barriers to entry.

To this point, the economic literature has presented only limited information regarding specific potential sources of entry barriers into the long-distance market. One area that has generated some amount of study has been the issue of long-distance consumers' willingness and ability to switch long-distance carriers. Clearly, the higher switching costs are and the lower the willingness of consumers to switch in response to asymmetries across carriers, the greater the degree of pricing latitude (power) incumbent carriers are likely to possess. The 'willingness to switch' issue has been an important topic in the literature in this area from the moment of the divestiture.

For example, Shepherd (1995) argues that if markets are perfectly well functioning (i.e., contestable), then price differentials by AT&T should have been unsustainable from the moment of the divestiture. Yet, in the early post-divestiture period, Shepherd finds that AT&T's prices were set at levels that were up to 30 percent higher than MCI or Sprint. This finding, then, is viewed as evidence that the long-distance market exhibits substantial entry barriers. Katz and Willig (1983), however, point out that the advent of equal access for all presubscribed carriers, together with the growing comfort on the part of consumers with alternative long-distance carriers, would lead to a situation in which consumers would increasingly be willing to switch carriers in response to observed price differentials. These price differentials, then, should be expected to narrow as this process unfolds. And, indeed, that is precisely what has happened.

A more detailed study of switching and search costs in the long-distance industry has been conducted by Knittel (1997, p. 533). Specifically, in the spirit of Klemperer (1987), Knittel argues that, to the degree a market is characterized by either switching or search costs, these characteristics may impair competitive performance. He then constructs an empirical model of the determinants of price-cost margins in the interexchange industry with proxies for the degree of switching and search costs in the industry. The price-cost margin is constructed using tariffed long-distance rates over the 1984–1993 period as the price variable, and the tariffed levels of access charges as a proxy for marginal cost[16]. Search costs are proxied by the level of advertising expenditures in the industry. According to Knittel, the larger are advertising expenditures, the greater the information at customers' disposal and the lower the resulting search costs. Additionally, the

[16] It is generally recognized that the use of basic tariffed rates to represent prices in this industry is biased due to the presence of discount programs. Moreover, that bias has grown over time as these programs have proliferated.

standard deviation of rates is used as a proxy variable for search costs. That is, the more widely dispersed are prices, the higher are search costs, *ceteris paribus*. Switching costs are proxied by the charge local exchange carriers place on long-distance firms when an order is placed to switch a customer from one long-distance carrier to another.

The empirical results indicate that advertising negatively influences price-cost margins, though at a diminishing rate. Knittel suggests that this finding indicates potential 'benefits to advertising' contradicting the standard view that advertising creates barriers to entry that raise 'market power.' The results also indicate a positive and statistically significant coefficient on the standard deviation of long-distance rates. Knittel's interpretation of this result is that search costs are an important determinant of price-cost margins in the long-distance market. An alternative explanation, however, immediately arises. As noted above, during the 1984–1993 period, local exchange companies were implementing equal access, which gradually removed the dialing disparity between AT&T and its rivals over time. This reduction in product differentiation predictably led to reductions in the standard deviation of long-distance rates across carriers. Moreover, the homogenization of long-distance services over this period is likely to have resulted in reduced barriers to entry and a predictable compression of price-cost margins[17]. Thus, the observed positive correlation between diminishing price differentials and declining price-cost margins may be due to this mutual correlation with the percent of lines converted to equal access.

Knittel's interpretation of switching costs as a major determinant of pricing in the interexchange industry is also subject to question. Specifically, the variable used to proxy for switching costs is the fee imposed by local exchange carriers on interexchange firms that have successfully secured a customer from a rival. This charge (typically $5.00) is, however, routinely absorbed by long-distance companies. That is, consumers more often than not never see such charges. Indeed, as pointed out by Knittel, there are often quite lucrative inducements offered to entice customers to switch (e.g., free long-distance calling, free airline miles, or cash) that create, in effect, 'negative switching costs.' Moreover, if switching costs were acting to enable elevated price-cost margins, then it should be the case that these carriers would be reluctant to waive such charges.

Thus, it is difficult to accept that the fees imposed on customers for changing long-distance carriers are acting as a real 'switching cost' for long-distance customers or that such charges are acting to support supra-competitive pricing. Indeed, industry data indicate that tens of millions of orders to switch long-distance carriers are made every year. This extraordinary level of switching among carriers would seem to provide a strong *prima facie* case that once reimbursements

[17] Kahai, Kaserman and Mayo (1995) find evidence that the homogenization of long-distance services that occurred with the implementation of equal access (which eliminated the dialing disparity between AT&T and its rivals) resulted in reduced barriers to entry and greater numbers of long-distance competitors.

to customers for switching are taken into account, the level of 'real' switching costs in the long-distance market is very low[18]. An alternative explanation for the finding that such fees are associated with elevated price-cost margins is that the measure of marginal cost used by Knittel does not include such fees. These fees, which are incurred by long-distance carriers, are, however, marginal to the acquisition of customers. The result is that the imposition of these charges is logically reflected in the prices being charged to long-distance customers.

Another potential source for barriers to entry into the long-distance industry is the presence of sunk costs. Hausman (1995, p. 365–366) has argued that barriers to entry into the long-distance industry are "extremely high" and that the source of these barriers is "the requirement of sunk costs to construct a new fiber-optic long-distance network." Without question, the deployment of a nationwide fiber optic network involves a large amount of (literally) sunk costs.

Two considerations suggest, however, that the presence of such high sunk costs associated with the construction of a nationwide fiber-optic network do not warrant the conclusion that barriers to entry into the long-distance market are necessarily high. First, the argument uses the wrong standard to judge the height of barriers to entry. Specifically, entry barriers should be measured by examining the economic characteristics of costs for the most likely mode of entry. That is, profit-maximizing firms will typically seek to enter markets via the least-cost mode of entry that minimizes their exposure to losses in the event that the new venture fails. In the case of the long-distance industry, this least cost path does not involve *de novo* construction of a fiber-optic transmission network but, rather, entry by leasing existing transmission capacity. That is, new entrants more often than not enter the industry as resellers – buying transmission capacity in the wholesale market and selling the long-distance services to retail customers[19]. As these resellers grow and expand their customer bases, a point is reached where it may become economically more attractive to own facilities rather than purchase capacity in the wholesale market. Indeed, once a customer base is in place, the prospect of a reseller's exit from the market diminishes, and the consequence is that the risk associated with the purchase of the sunk cost transmission facilities is correspondingly mitigated. In short, the feasibility of resale entry is likely to reduce the otherwise entry-deterring properties of high sunk cost facilities-based entry.

A second question regarding Hausman's claim of "extremely high" barriers to entry into the long-distance market centers around the fact that hundreds of firms have, in fact, entered that market in the post-divestiture market (See Figure 1), albeit, most as resellers. Nonetheless, successful entry by this many firms seems to

[18] See Kaserman and Mayo (1996).

[19] Resellers are simply non-vertically-integrated carriers. The competitive discipline these firms are able to exert, then, depends upon the competitiveness of the upstream market for transmission capacity. That competitiveness, in turn, has been enhanced by the emergence of so-called "carriers' carriers" that lease transmission facilities, but do not market final services to consumers. In addition, some private networks have leased transmission capacity to resellers on occasion.

provide a compelling case that, in contrast to Hausman's claim, barriers to entry into the long-distance market are low.

Another potentially important structural determinant of the degree to which the market is likely to generate competitive performance centers on the extent to which incumbent firms hold excess capacity. In particular, industrial organization models of industry behavior suggest that the degree of excess capacity will affect the ability of the industry to sustain supra-competitive pricing. Moreover, Staiger and Wolak (1992) refer to a "large body of empirical evidence" that supports the proposition that the incentives for vigorous price competition are heightened by low capacity utilization[20]. The logic of this argument is straightforward. Where there is excess capacity, the marginal cost of increasing the individual firm's output is reduced, and the incentive to expand output at any given price is heightened correspondingly.

In this regard, there appears to be considerable agreement that rapid deployment of fiber optic transmission facilities for long-distance services, together with marked advances in the electronic facilities on the 'ends' of the fiber, have drastically expanded the installed capacity of long-distance carriers. Several studies of the capacity of long-distance carriers indicate that the industry has been marked by substantial excess capacity[21]. Such capacity tends to serve two pro-competitive purposes. First, under the Staiger and Wolak (1992) logic, such capacity should tend to promote more competitive pricing by the firms holding that capacity. Second, excess transmission capacity should enhance the bargaining position of resellers. Both effects, of course, encourage more competitive outcomes in the final product market.

We conclude our discussion of the structure of the long-distance market by noting that, over the past decade, an argument has arisen (e.g., MacAvoy [1995], Hausman [1995], Taylor and Zona [1997]) that long-distance firms are engaged in tacit collusion to keep prices above competitive levels. Chamberlin (1933) first developed the concept of tacit collusion. The basic idea is that, under certain conditions, rival firms in a highly concentrated industry may gravitate toward the joint profit-maximizing (i.e., monopoly) price and output without actually entering into an explicit agreement to fix prices. Whether this sort of behavior is likely to occur, however, is highly dependent upon the specific structural characteristics of the market in question. For tacit collusion to arise, the industry structure must be conducive to a 'meeting of the minds' that must occur to sustain this sort of coordinated conduct. Thus, an examination of structural characteristics of the market provides indirect evidence on the prospects that tacit collusion will arise.

The above examination of the structural characteristics of the long-distance market suggests that this market is not conducive to tacit collusion[22]. While the

[20] But, note Spence (1977), who argues that excess capacity may be used to create entry barriers.

[21] See e.g., the discussion in FCC (1995), MacAvoy (1996, pp. 93ff.), and AT&T Bell Laboratories (1995).

[22] This conclusion is also consistent with the findings of Ward (1999).

industry's output is concentrated, which increases the prospects for such tacit collusion, *ceteris paribus*, there are a host of other structural characteristics of the market that act to diminish the likelihood of tacit collusion. These structural characteristics, which are described in detail in Kaserman and Mayo (1996), include, *inter alia*: (1) the absence of significant barriers to entry; (2) the presence of excess capacity; (3) asymmetries of market shares among the largest competitors; (4) the widespread use of highly tailored pricing plans and promotion activities across the various long-distance carriers; (5) the rapid technological change in the provision of long-distance service; (6) the skewness of market demand; and, (7) the very large number of providers of long-distance service operating in the market.

On balance, the host of structural characteristics present in the long-distance market that act to deter the emergence of tacit collusion are likely to overwhelm the simple fact that the industry is highly concentrated. Finally, we note that, while market structure provides (or in this case fails to provide) the preconditions for tacit collusion, it is possible to conduct more direct empirical tests of firm behavior to determine whether tacit collusion is, in fact, present in a market. We discuss and present such a direct empirical test in a subsequent section of this chapter.

2.2. Conduct

Within the S-C-P paradigm, industrial organization economists have posited that several aspects of industry conduct flow from industry structure. In the long-distance market, this relationship has been explored most fully by MacAvoy (1995, 1996). Specifically, he argues that, at the time of the divestiture, there were considerable incentives on the part of AT&T's rivals to price aggressively in order to attract market share from AT&T. The fundamental economic source of this aggressiveness came, according to MacAvoy, from two market characteristics. First, long-distance prices were set by regulation at artificially high levels. These high rate levels were the product of the evolution of a long-distance-to-local-exchange subsidy that developed over many years within the unified Bell System prior to divestiture. Moreover, prior to 1989, rate-of-return regulation constrained AT&T's ability to lower these rates in response to the competitive incursions of its rivals. Second, new entrants (who did not enjoy the benefits of equal access) received deep (55 percent) discounts on the purchase of 'access' from local exchange carriers. The result, MacAvoy argues, provided both the incentive and opportunity for AT&T's rivals to rapidly capture market share.

MacAvoy goes on to argue that, by 1989, both the regulation-mandated high prices and the cost disparity between AT&T and its rivals had ended. The result, he suggests, was the emergence of a much less aggressive – more cooperative – pricing regime in the long-distance marketplace. To test the proposition that market conduct was evolving to tacit collusion, MacAvoy conducts a simple

econometric exercise. Specifically, for several individual long-distance services, he develops an estimate of a price-cost margin (PCM). The marginal cost of providing long-distance services is taken to be equal to the sum of access charges paid by long-distance carriers for originating and terminating access, plus the incremental operating costs of transporting the calls once they emerge from local exchange carriers' networks. Drawing upon an unpublished study, he proxies the latter costs by an estimate of $0.01 per minute which, apparently, is assumed to remain constant over both time and carriers. The calculated price is a weighted average across AT&T, MCI, and Sprint of the tariffed prices for seven long-distance services.

The resulting price-cost margin series is then regressed on a calculation of the level of the industry's HHI. The equation estimated by MacAvoy is given by:

$$\frac{P - MC}{P} = \alpha + \beta_1 HHI + \sum_i \beta_i S_i, \tag{1}$$

where the dependent variable is a proxy for the economic price-cost margin; *HHI* is the Herfindahl-Hirschman Index calculated for the largest three carriers only; and, S is a vector of six service-specific dummy variables used to indicate particular services (e.g., switched outbound, and dedicated outbound). The above equation is estimated separately over two sample periods. First, the model is estimated for the 1987–1990 period. And, given the hypothesized shift in firm incentives to compete aggressively that occurred in 1989, the model is estimated again for the 1991–1993 period.

The estimation results reveal a negative and statistically significant coefficient on the variable of principal interest (i.e., β_1) for both periods. Moreover, the *HHI* coefficient resulting from the estimation over the latter time period yields a parameter estimate that is larger in absolute value than the estimate obtained from the earlier time period. MacAvoy interprets these results as evidence of emerging tacit collusion among AT&T, MCI and Sprint. That is, he argues that the negative correlation between the observed price-cost margin and the degree of concentration in the market is inconsistent with the presence of effective competition.

While MacAvoy's interpretation of these results provides the provocative conclusion that the long-distance industry is characterized by tacit collusion, the model and methodology upon which this conclusion rests are open to challenge. Specifically, while the theoretical construct of the Lerner Index (the true measure of market power) is defined by conceptually accurate measures of price and marginal cost, direct empirical estimation of these constructs is notoriously inexact. Indeed, it is precisely because of the difficulty likely to be encountered in measuring marginal cost that it is generally conceded that direct estimation of the Lerner Index is not generally feasible[23].

[23] See Landes and Posner (1981).

This difficulty in measuring marginal cost raises considerable questions about the measurement of the dependent variable. For example, while MacAvoy estimates the non-access portion of incremental costs to be $.01 per minute, Crandall and Waverman (1995) estimate these same costs to be between $.03 and $.08 per minute. Thus, the cost component of the calculated price-cost margin used as the dependent variable must be viewed cautiously.

Moreover, long-distance services provide an equally daunting challenge with respect to the measurement of 'price' in any direct calculation of price-cost margins. Specifically, measurement of the price of long-distance service is made more difficult because: (1) the amount any given customer pays for long-distance service has, over time, increasingly reflected a variety of discounts; and (2) complex tariff structures that include a 'fixed' component and which may also vary by distance, duration, time of day, and so on, render a single price figure ambiguous. Thus, use of some calculated price-cost margin as a measure of the degree of market power is, at best, only a crude proxy; and at worst, it is a highly misleading indicator.

A second potential source of concern regarding the MacAvoy model is the implication that observed service-specific price-cost margins should be related only to the overall degree of concentration in the market. That is, the MacAvoy model is constructed in such a way that the only variable that is allowed to drive inter-temporal movements in the calculated price-cost margin is changes in the HHI. All other variables incorporated in the model are service-specific dummy variables which do not vary over time. Accordingly, the coefficient on the HHI variable is, by necessity, the only one capable of picking up the effects of any omitted variables that do vary with time. For example, to the extent that the market was expanding rapidly during this period (as indicated in Figure 2), then price-cost margins over the period may have experienced upward movement as a result of simple demand-side pressures. In addition, other factors such as input prices (e.g., access charges), technological change, increasing product homogeneity, and so on, are likely to have affected price-cost margins over this period. Consequently, it must be seen as a considerable stretch to infer tacit collusion from the mere presence of rising price-cost margins and falling market concentration.

A third fundamental problem with the model estimated by MacAvoy is that it suffers from several serious econometric problems. First, economic theory suggests that the Lerner Index is related not only to market concentration, but also to entry conditions, and other factors such as those mentioned above. The MacAvoy model omits any measure of any of these determinants of performance. Thus, the model is likely to suffer from omitted variable bias to the extent that these factors have varied over the sample period. Also, economic theory indicates that the degree of market concentration is likely to be determined in part by price-cost margins. Thus, the principal (and only continuous) right- hand-side variable is likely to be endogenously determined. Finally, MacAvoy's finding that $\beta_1 < 0$ – in

both periods – is a theoretically anomalous result. No standard market model predicts that declining concentration should lead to an increase in price-cost margins. To ascribe this empirical result to tacit collusion seems a bit of a stretch. It is simply not clear why an observed negative relationship between the HHI and price-cost margins represents unambiguous empirical evidence of tacit collusion. Accordingly, a more fully specified system of equations is likely to be necessary before reaching confident conclusions regarding a breakdown of competition on the part of firms in the long-distance industry.

Another empirical study of the conduct of long-distance carriers focuses on concerns that, as a 'dominant' provider, AT&T might engage in predatory practices in a less regulated environment. To investigate this issue, Kahai, Kaserman and Mayo (1995) propose an empirical test for identifying predatory behavior. Unlike tests that focus principally on the relationship of prices to some measure of costs (e.g., average variable costs under the well-known Areeda-Turner (1975) test), these authors focus on whether relaxed regulation of AT&T has resulted in the exit of long-distance carriers after accounting for various other determinants of the number of carriers in a given state. The underlying logic of this approach is that, if predation is occurring, then, *ceteris paribus*, some exit should be observed. Thus, unlike predation tests that focus on pricing, the methodology proposed by Kahai, Kaserman and Mayo provides for a more comprehensive test that includes essentially any predatory action (e.g. preemptive investments, denial of necessary interconnections, etc.) that might result in the exit of rivals.

The empirical model employed to implement this test draws upon the fact that, in the wake of the divestiture, different states adopted different regulatory approaches toward AT&T[24]. Thus, after accounting for other factors that influence the number of firms in the market, a necessary condition for the presence of successful predation is the exit of firms[25]. The data used to perform the test include, as the dependent variable, the number of long-distance firms operating in a state in 1990 and 1992. Data for 40 multi-LATA states were gathered and the

[24] Indeed, this disparate treatment by the states provides a natural setting for testing the competitiveness of the long-distance market. Specifically, firm and market behavior in response to deregulatory measures will differ markedly depending on the state of competition in the market. Thus, after accounting for other determinants of market conduct, it is possible to utilize observed market behavior in response to regulatory changes to make inferences about the state of competition in the market. We return to this point below in Section 3.

[25] Kahai, Kaserman and Mayo point out several caveats to this conclusion. First, because the test relies upon observed exit, the applicability of the test is limited to cases where predation may have already occurred. That is, it is an ex post test that cannot be used as an ex ante detection mechanism. Second, any such test that relies upon observed exit cannot provide a test for punitive pricing that is undertaken to discipline rivals into adopting a more accommodating mode of behavior. And finally, if exit does occur, the test is not rich enough to determine the causes of that exit. That is, observed exit may be the consequence of aggressive competitive behavior or predation. Rather, the test is one-sided. It relies upon the fact that observed entry or the lack of exit is inconsistent with predation.

number of competitors was specified to be a function of the number of carriers in 1986, the size of the intrastate toll market, the relative importance of business customers in the state, the price-cost margin, the extent of equal access, whether the intraLATA intrastate toll market was open to competition and, finally, whether AT&T enjoyed deregulation. If, after controlling for the other determinants of the number of competitors, the deregulation of AT&T led to a decline in the number of carriers then it is possible that predation was the underlying source. Alternatively, if no exit of firms is attributable to the deregulation of AT&T, then predation would seem to be ruled out.

The empirical results indicate that, while the number of long-distance carriers operating in 1992 in a given state is statistically (and positive) influenced by the number of carriers operating in 1986, the size of the market, the share of the state's lines that are dedicated to businesses, and the degree to which states have implemented equal access, there has been no independent effect of relaxed regulation of AT&T on the number of long-distance carriers. Thus, the empirical test rejects the argument that AT&T has engaged in predatory behavior in response to relaxed regulation.

2.3. Performance

The divesture of the Bell Operating Companies from AT&T in 1984 gave rise to a number of concerns regarding the subsequent performance of the long-distance market. Specific performance characteristics of concern were, and are, the degree of allocative and productive efficiencies obtained, product quality and innovation in the market, and so on. On these dimensions, the industry has changed dramatically in the post-divestiture period. While there is an ongoing debate regarding the source of price changes in the provision of long-distance service (we discuss this controversy in Section 4 below), there is no question that the prices of long-distance services have fallen sharply in the post-divestiture period.

Figure 5 provides a comparison of changes in the CPI for all goods and services and changes in the CPI for interstate toll services[26]. There, we see that, in fourteen of the past seventeen years for which price data exist, the price of long-distance service has increased less rapidly than the overall inflation rate, and in twelve of those years, the nominal price of long-distance service declined. The tariffed rates (using AT&T as representative) have fallen by about 30 percent for directly dialed residential customers and by 20 percent for business customers over the 1984–1998 period[27]. The distribution of these price changes has varied depending upon the distance of the long-distance call. Specifically, relatively short-haul long-distance calls have experienced price increases, while calls of longer distance have received large price reductions[28]. This re-calibration of prices by distance is

[26] Tables 13.2 and 13.3 of FCC "Trends in Telephone Service," September 1999.
[27] FCC (1999), p. 14–1.
[28] FCC (1999), Table 14.4.

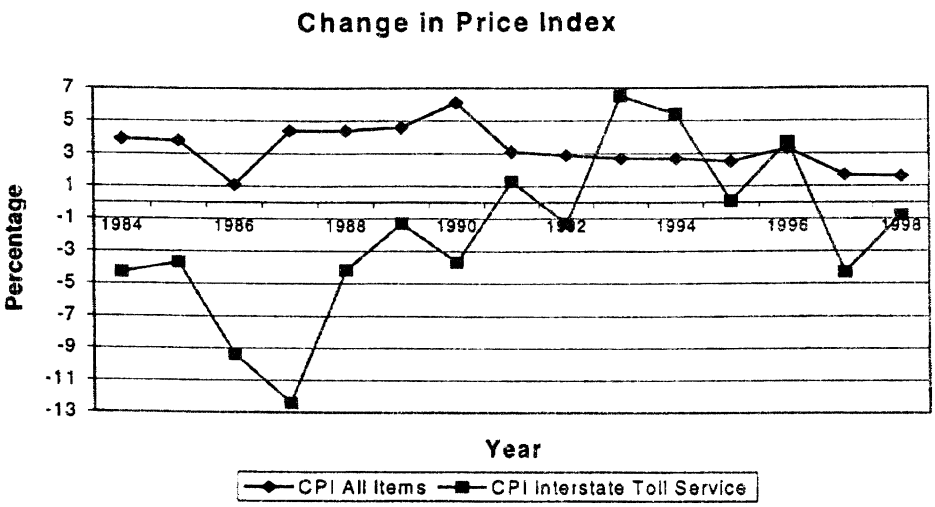


Fig. 5. Herfindahl Hirschman index.

consistent with competitive adjustments in the wake of the divestiture of AT&T. Specifically, prior to the divestiture (and for as long as AT&T remained rate regulated), the price of short-haul interLATA calls was held down (often below costs) while the price of long-haul toll calls was held artificially high. Thus, as would be expected in this situation, the emergence of competition and deregulation led to price increases on short-haul toll calls while the price of long-haul calls fell sharply. Table 1 displays the tariffed (undiscounted) rates in January, 1984 and 2000, for a residential five-minute daytime call using AT&T.

While tariffed rates for long-distance services have fallen in most years since the divestiture, the dramatic emergence and proliferation of discounted rates and calling plans available to long-distance consumers characterizes industry performance. In 1984, AT&T initiated its Reach Out America calling plan which afforded customers the opportunity, on a voluntary basis, to avail themselves of long-distance charges that were either structured as two-part tariffs, tapered rates, or block-of-time purchases[29]. Then, beginning with MCI's Friends and Family, which provided discounted calling to a 'calling circle' of the subscribers' 'friends and family,' the industry has witnessed an explosion in the number of discounted calling plans. Today, discounted plans are available from all carriers and incorporate a number of auxiliary benefits such as accelerated credit toward free air travel or trips to Disney World[30].

[29] For a discussion, see Mitchell and Vogelsang (1991).
[30] See, e.g., Andrews (1995, p. 48).

Table 1
Tariffed 5-minutes daytime call rates for AT&T

Mileage band of call	January 1984	January 2000	Percentage change
1–10	$.96	$1.30	35.4
11–22	1.28	1.30	1.6
23–55	1.60	1.30	−18.8
56–124	2.05	1.30	−36.6
125–292	2.14	1.30	−39.3
293–430	2.27	1.30	−42.7
431–925	2.34	1.30	−44.4
926–1910	2.40	1.30	−45.8
1911–3000	2.70	1.30	−51.9

Yet another performance characteristic of the long-distance marketplace is quality. Relatively little formal economic attention has been paid to this characteristic. Taylor and Zona (1997) provide evidence taken from AT&T that overall blockage rates have fallen over time and are very comparable for geographic areas that have equal access and those that do not. Nonetheless, Taylor and Zona conclude that "facilities-based competition is not the driving force behind improvements in quality."

2.4. Summary

What lessons can be gleaned from the analysis of structure, conduct and performance over the past fifteen years? Unfortunately, relatively little consensus has emerged. Several authors have concluded that industry structure, conduct and performance of the long-distance industry have moved steadily toward effective competition in the wake of the 1984 divestiture of AT&T. This conclusion is, however, not without dissent. It is for that reason that it is necessary to turn to other avenues by which to shed light on the central questions surrounding the extent of competition in the long-distance industry. Accordingly, we now turn to those other avenues.

3. Empirical analysis of relaxed regulation

A fundamental premise of the divestiture of the U.S. telecommunications industry in 1984 was that structural separation would substantially undermine, if not completely eliminate, the market power held by AT&T. The reason for this belief was the determination that the source of AT&T's market power was its control over local telephone exchange and access facilities. Without control over 'the last mile' of the network, AT&T would lose its ability to control the price, quality,

terms or conditions of the access facilities that its downstream (long-distance) rivals need to compete. Judge Greene, who oversaw the divestiture, wrote,

> "Once AT&T is divested of the local Operating Companies...it will be unable to subsidize the prices of interexchange services with revenues from local exchange service or to shift costs to competitive interexchange services." Thus, the Court concluded, "With the removal of these barriers to competition, AT&T should be unable to engage in monopoly pricing in any market"[31].

Despite Judge Greene's optimistic outlook for competition in the long-distance marketplace, the act of divestiture did not immediately result in the deregulation of the long-distance industry. Instead, AT&T simply inherited the same rate-of-return regulatory structure that it had faced prior to the divestiture. While there was some initial discussion of deregulation of AT&T, the policy path quickly evolved to a more "studied" approach[32]. Indeed, it was not until 1989 that the FCC eliminated rate-of-return regulation of AT&T's interstate services[33]. At that time, the FCC implemented a price cap regime that remained in place until 1995, when it finally ended direct price regulation in the industry[34].

State-level regulatory policies also evolved over this period. In general, several states moved much more quickly and aggressively than the FCC to eliminate traditional rate-of-return regulation. For example, in mid-1984, Virginia's public utility commission announced that it was eliminating rate-of-return regulation and granting AT&T full pricing flexibility on its intrastate services. Subsequent to Virginia's deregulation, a number of other states followed suit. At the same time, other states moved to relax regulatory controls by implementing price caps on AT&T's services[35]. Still others retained rate-of-return regulation for a number of years.

Industrial organization economists quickly recognized that the state-level variation in the degree to which AT&T was price regulated provided a fertile field for empirical investigation of the competitiveness of the long-distance market. Specifically, if AT&T retained its pre-divestiture market power, then, assuming that the regulatory controls were binding, relaxed regulation of its services might be

[31] United States v. AT&T, 48 PUR 4th 227, F. Supp. at 172 (D.D.C. 1982)

[32] For examples of early calls for the deregulation of the long-distance industry, see Katz and Willig (1983) and Kaserman and Mayo (1986).

[33] Regulatory jurisdiction for long-distance services is bifurcated in the United States. Interstate long-distance services fall under the jurisdiction of the Federal Communications Commission, while intrastate telecommunications fall within the bailiwick of state public utility commissions.

[34] In re Motion of AT&T Corp. to be Reclassified a Non-Dominant Carrier, Order, CC Docket No. 79-252, FCC 95-427, October 23, 1995. Despite this deregulation, however, the FCC still manages to exercise some indirect controls over AT&T's pricing decisions. For example, in agreeing to approve the recent CALLS proposal to reduce interstate access charges, the FCC required AT&T to drop its minimum monthly charges.

[35] For a discussion of the price caps mechanism applied to AT&T, see Mitchell and Vogelsang (1991). Also, see Kaserman and Mayo (1990) for a discussion of state-level regulatory changes.

expected to lead to price increases. Alternatively, if regulation was serving to restrain AT&T's ability to respond aggressively to its new rivals through price reductions, then relaxed regulation should result in price declines.

The first such study to emerge in this vein was by Mathios and Rogers (1989). They utilized state-level data from 1983–1987 to examine the effects of state-to-state variations in regulatory policies toward AT&T on intrastate interLATA long-distance prices. Price data were collected from tariffed rates for ten mileage bands for 39 states[36]. Using these tariff rates, they constructed a set of prices for a five-minute daytime long-distance call within each state for a specified set of mileage bands. Of the 39 states in their data set, 28 states had instituted some form of relaxed regulation between 1984 and 1987.

Given these data, Mathios and Rogers then constructed a model of AT&T's prices of long-distance services. Specifically, long-distance prices were modeled as a function of a standardizing set of demand-side variables, e.g., the state's population, supply-side characteristics, e.g., the average wage rate of telecommunications employees in the state, and state-level variations in the regulatory regime. Mathios and Rogers include a dummy variable to account for whether particular states had reduced access charges assessed on long-distance carriers over the post-divestiture period. The empirical results indicate that "the price of a five minute call, on average, is 7.2 percent lower in states that allow pricing flexibility." This result suggests that AT&T did not possess substantial market power in those states where it obtained reduced regulation.

While the Mathios and Rogers paper provides the seminal research on the implications of relaxed regulation on pricing in the long-distance market, several questions have plagued the study. For instance, given the prominence of access charges as a cost item in the provision of long-distance services, the use of a simple dummy variable to indicate whether a given state reduced access charges between 1984 and 1987 seems unnecessarily crude. Similarly, the principal results of the Mathios and Rogers study center around the impact of a state establishing *any* pricing flexibility. Thus, the Mathios and Rogers study is unable to differentiate between varying degrees of relaxed regulation of prices, e.g., price caps. Finally, as pointed out by Crandall and Waverman (1995), the Mathios and Rogers model does not account for structural market factors that might account for variations in the degree of competitive pressure faced by AT&T in different states. Despite these limitations, however, the Mathios and Rogers' results provide some relatively early econometric corroboration of the S-C-P results that seem to suggest that once AT&T was allowed pricing flexibility, it took advantage of its newfound freedom to reduce rates in response to the competitive pressures it faced. This evidence is suggestive of a lack of substantial market power by AT&T in the long-distance market as early as 1987.

[36] Data for intrastate long-distance services provided by AT&T were not available in 11 states which had only a single LATA.

Kaestner and Kahn (1990) performed another study of the effects of regulatory variations on the prices of long-distance service. These authors begin by positing a general model of price-cost markup in an oligopolistic setting:

$$P = MC[1/(1 - (S(1 + Q)/N)], \tag{2}$$

where P is AT&T's price of long-distance service in a given state, MC is the marginal cost of providing long-distance service, S is AT&T's share of the market, Q is a conjectural variation term to represent AT&T's beliefs regarding the response by its rivals to changes in its output, and N is the absolute value of the industry price elasticity of demand. The natural logarithm of Equation (2) yields a generic formula:

$$\ln(P) = \ln(MC) + \ln(markup) \tag{3}$$

where the markup is:

$$\frac{1}{1 - S(1 + Q/N)}. \tag{4}$$

Kaestner and Kahn then posit that the MC of providing long-distance service is a function of the level of intrastate access charges ($ACCESS$), the type of regulation (REG), and state specific cost factors ($STATECOST$). The markup of prices is modeled as a function of the type of regulation and the market share of AT&T's rivals ($SHARE$). After introducing demand considerations ($STATEDEMAND$), a final reduced form is generated:

$$\begin{aligned} lnP = {} & \beta_0 + \beta_1 ACCESS + \beta_2 REG + \beta_3 SHARE + \beta_4 STATECOST \\ & + \beta_5 STATEDEMAND + \varepsilon. \end{aligned} \tag{5}$$

Because of the potential for AT&T's prices to affect the value of SHARE, Kaestner and Kahn specify a SHARE equation:

$$SHARE = f(ACCESS, REG, PRICE, EACONV, DENSITY) \tag{6}$$

where $EACONV$ is the percentage of the state's lines converted to equal access and $DENSITY$ is the population per LATA within the state. Together, Equations (5) and (6) form a simultaneous system that Kaestner and Kahn estimate via three-stage least squares for state-level data collected over the 1986–1988 period.

The estimation yields several interesting findings. First, both relaxed regulation and deregulation of state-level long-distance pricing lead to significant price reductions. Unfortunately, the empirical model is not sufficiently specified to yield disaggregated estimates of the effects of deregulation on the cost of producing

long-distance services and on the markup of prices above costs. The aggregate results, though, clearly indicate that the net impact of these two influences was to reduce prices in the wake of deregulatory policies. This basic finding, then, tends to corroborate the Mathios and Rogers (1989) results.

Second, the estimation reveals that *SHARE* is highly sensitive to changes in the price of intrastate long-distance service. Specifically, the empirical results indicate that, for a ten-cent increase in the price of a five-minute AT&T daytime call, competitors' collective market share grows by 3.3 percent. This finding suggests that AT&T's competitors are both willing and able to expropriate share from AT&T if they find that prices are high, which, in turn, supports the view that entry barriers are low in this industry.

Third, Kaestner and Kahn find that AT&T's price is highly sensitive to changes in the price of carrier access. A ten-cent reduction in the price of access for a five-minute daytime call was found to lead to a reduction in the corresponding price over the 1986–1988 period of 28 cents[37]. Finally, while finding that relaxed regulation led directly to lower prices, the estimations fail to yield a significant negative impact of competitors' share on AT&T's prices. This result is interpreted by Kaestner and Kahn as suggesting that increased competition did not lead to lower prices. This finding, however, is subject to an alternative interpretation. Specifically, the impact of competition in the intrastate markets under the Kaestner and Kahn interpretation is proxied solely by the extent of market share captured by AT&T's rivals. It is well known, however, that market share is only one of a set of variables that determine the extent of competitive pressure exhibited in a marketplace[38]. Indeed, to the extent that the long-distance market is characterized by low barriers to entry, one would not expect to find a significant effect of changes in market share on price. Moreover, Kaestner and Kahn's interpretation appears to be at odds with their principal finding that reduced regulation of AT&T led to lower prices. Thus, these authors' interpretation of this last finding remains ambiguous.

In conclusion, the state-level variation in the extent to which AT&T was given pricing flexibility in the wake of the divestiture created a natural experiment whereby it is possible to test for the underlying market power enjoyed by AT&T after 1984. Specifically, rather than estimating the degree of market power from a structure – conduct – performance perspective, these studies allow indirect evidence on AT&T's market power by examining how its behavior changed in response to changes in the severity of the pricing regulation it faced. Overall, the studies that we have reviewed indicate reduced regulation of AT&T at the state level has tended to result in lower prices for long-distance services, *ceteris paribus*. Accordingly, the results from these studies suggest that the divestiture of AT&T

[37] The relationship between access charge changes and changes in the price of long-distance service is discussed in detail in Section 4 below.

[38] See, e.g., Landes and Posner (1981) who caution against solely focusing on market share characteristics in making assessments of market power.

from the local exchange facilities that were hypothesized to be the source of its pre-divestiture market power has, in fact, accomplished the job of substantially dissipating AT&T's market power.

4. The access charge pass through debate

A third methodological approach that has been used to assess the intensity of competition in the long-distance market has relied upon comparisons of long-distance carriers' price, or revenue, changes and contemporaneous changes in the rates local exchange companies charge, or payments they receive, for carrier access services[39]. Since divestiture, such services have constituted the single most important input, in terms of its share of overall costs, employed by long-distance companies in providing service to final consumers. These access charges have generally comprised between 30 and 60 percent of long-distance carriers' total costs, depending upon the time period and whether the non-equal access discount was in force[40]. These access rates have been reduced substantially over time at both the state and federal levels since divestiture. For example, interstate access charges have fallen from over 17 cents per minute in 1984 to about 3 cents per minute today[41].

Given these access charge reductions, several authors have argued that the observed performance of long-distance carriers in passing through these input price changes to their customers through output price reductions provides direct evidence of the effectiveness of competition in the long-distance market. Specifically, any failure to fully pass through access charge reductions has been interpreted as evidence of non-competitive performance[42]. These arguments and the calculations used to support them have sparked considerable debate in the literature on this subject.

The pass through controversy became particularly prominent in regulatory deliberations regarding the likely benefits of allowing the Regional Bell Operating Companies (RBOCs) to enter in-region interLATA long-distance markets under the provisions of Section 271 of the 1996 Act. Specifically, the RBOCs argued that AT&T and other long-distance carriers had failed to fully pass through their prior

[39] These inter-company access charge payments were designed to replace the largely intra-firm system known as Separations and Settlements that was used to allocate costs and transfer revenues across jurisdictional boundaries prior to divestiture.

[40] After divestiture, local exchange companies gradually upgraded central office equipment in order to provide access services to non-AT&T long-distance firms that was equal in quality to that being provided to AT&T. Where these firms continued to receive non-equal access, they were awarded a 55 percent discount below the access rates paid by AT&T. These discounts were phased out as equal access was implemented.

[41] State-level access charges (which apply to all intrastate toll calls) have generally tended to be above interstate levels and have varied widely across the states. All, however, exceed the marginal cost of providing the access service, which has been estimated to be on the order of 0.5 cents per minute or less.

[42] See, e.g., Taylor and Taylor (1993) and Taylor and Zona (1997).

access charge reductions. That failure, in turn, was alleged to provide evidence that the long-distance market was not performing competitively. As a result, these parties argued that the expected benefits of allowing the RBOCs to enter the interLATA market would be substantial as overall performance would improve significantly. In addition, the RBOCs have used similar pass-through arguments to oppose regulatory reductions in the level of access charges, holding that any such reductions would fail to benefit consumers as they are usurped by the long-distance carriers.

At the most fundamental level, flow-through arguments are grounded upon two essential propositions—one theoretical and one empirical. These are:

Proposition 1: (Theoretical) *Valid inferences regarding the intensity of competition can be made from observations of firms' access charge flow through behavior; and*
Proposition 2: (Empirical) *AT&T and other long-distance firms have not flowed through access charge reductions fully to consumers.*

Consider each of these propositions in turn. First, as a theoretical matter, is it possible to infer the presence or absence of competition from an analysis of whether and to what extent cost changes experienced by firms in an industry, from any source, are reflected in output price changes? In particular, if an industry's price changes by less than individual firms' marginal costs change, can we conclude that the industry is not performing competitively? Second, setting aside the question of theoretical validity, as an empirical matter, have AT&T and other long-distance carriers reduced their output prices to fully reflect the cost reductions experienced as a result of access charge rate changes? The following subsections consider each of these questions in turn.

4.1. Theoretical issues

The general approach of using observed pass through performance to draw inferences regarding the competitiveness of a market was first introduced by Sumner (1981) in a study of the U.S. cigarette industry. In this study, Sumner notes that profit maximization by individual firms requires that

$$P_i = [\eta_i/(\eta_i + 1)]MC_i, \tag{7}$$

where P_i is output price, MC_i is marginal cost, and η_i is the price elasticity of the individual firm's demand. In addition, he explains that marginal costs are determined by a set of exogenous factors that may vary across firms, X_i, and by a common tax rate, R, which is constant across firms (but varies across states and/or time). Thus, Sumner specifies

$$MC_i = X_i\beta + R + \mu_i, \tag{8}$$

where μ_i is a random error term.

Aggregating across firms and combining Equations (7) and (8), Sumner obtains an estimable equation relating output price to both X and R, where the coefficient attached to the tax rate variable is

$$\theta = \frac{\eta}{\eta + 1}, \tag{9}$$

where $\eta = \sum_{i=1}^{n} S_i \eta_i$ is the market share weighted average of the individual firms' price elasticities of demand. As such, η is a measure of the average intensity of competition faced by the firms in this industry. Given an estimate of θ, η is easily calculated from (9).

The question here is whether this approach, which seeks to infer the degree of competition in a market by the association of price changes to cost changes, is, in general, a theoretically valid method for gauging the intensity of competition in a market. That is, how robust is the basic logic underlying this method to changes in the specific assumptions, both explicit and implicit, contained in Sumner's analysis? As it turns out, the answer is: not very. The applicability of this approach depends critically upon a number of assumptions that may not necessarily hold in a given application. Several of these are discussed in the paragraphs that follow.

Constant elasticity demand: In a comment on Sumner's (1981) paper, Bulow and Pfleiderer (1983) show that the relationship between the responsiveness of output price to a change in marginal cost and the price elasticity of individual firm demand is quite sensitive to the particular functional form exhibited by that demand. Specifically, these authors demonstrate that:

$$\partial P / \partial MC_i = \eta_i / (\eta_i + 1) \tag{10}$$

holds *only* in the case of a constant elasticity demand curve. For all other functional forms, including linear, the fraction of cost changes passed on in the final output price will not equal the above expression regardless of the intensity of competition in the relevant market[43]. Indeed, Bulow and Pfleiderer (1983) conclude that: "Only when the elasticity of demand is constant is $p'(c)$ closely related to η." [Footnote omitted.] Thus, the ability to interpret observed pass through performance as an indicator of the competitiveness of a market is limited to cases in which the constant elasticity of demand assumption is met.

Constant marginal costs: A second limitation of Sumner's (1981) approach arises from the assumption of constant marginal costs. Where an individual firm's marginal costs rise with increasing output, or where the industry's marginal costs increase with entry, any input price changes cause not only a shift in costs, but also a movement along the relevant cost curve as the market equilibrates. In that case,

[43] The reason for this result is apparent from inspection of equation (7), above. Where η_i is not constant, a change in MC_i will lead to a change in the equilibrium η_i as well. As a result, Equation (10) will not hold.

then, the impact of an input price change will be spread across both consumers and producers depending upon the relative elasticities of supply and demand, just as the incidence of a tax is spread across these groups. As a result, with non-constant costs, input price changes will not be fully reflected in the corresponding output price[44].

In the long-distance industry, it is frequently assumed that marginal costs are constant, inasmuch as a substantial portion of such costs consist of carrier access charges which do not vary as output is increased. While such an assumption may represent a reasonable approximation for small output variations for certain services that rely upon switched access, it appears less likely to hold for larger variations in output in the long run across all services. Indeed, to the extent this industry is not characterized by natural monopoly conditions, increasing marginal costs appear more likely to hold.

Single product: Sumner's (1981) analysis was applied to a single-product industry – cigarettes. Suppose, however, that the relevant market consists of multi-product firms[45]. When such firms experience an input price change, the extent to which that change will be reflected in the individual products' prices will depend, *inter alia*, upon the input's cost shares for these products and the set of demand elasticities, and cross elasticities, that apply to them. Clearly, not all output prices will decrease by the same amount, even under competitive market conditions.

In the telecommunications industry, carrier access services comprise different cost shares across the individual products supplied. Indeed, some services, such as private lines, do not even employ local exchange switched access services as an input[46]. Also, price elasticities of demand vary across individual services as well[47]. Therefore, it would not be valid, in this setting, to examine pass through performance on a single product, or even a subset of products. As Beard (1999) explains, the relevant competitive standard in this case focuses on the impact on *suppliers*, not customers. Specifically, competition requires that an exogenous input price reduction does not result in windfall profits for producers. Consequently, the analysis must focus on the overall impact on sellers, not on the prices of individual products to buyers.

[44] This result, too, may be seen by inspection of Equation (7), above. If $MC_i = MC_i(q_i, A)$, where q_i is the firm's output and A is the price of access, then $\partial MC_i/\partial A = (\partial MC_i/\partial q_i)(\partial q_i/\partial A) + \partial MC_i/\partial A$. Consequently, Equation (10) will not hold in this case as well.

[45] Beard (1999) provides a theoretical analysis of the pass through issue in a multiproduct setting. Our comments here draw, in large measure, from that analysis.

[46] Specifically, customers can gain access to long-distance networks using either 'switched' access or 'special' access. Switched access is typically used by smaller and medium-sized users and is routed through the switches of the local exchange service provider for that customer. Special access is used by larger customers and involves a direct link between the customer and the interexchange carrier, without the use of the local exchange company switch. The prices of switched and special access are markedly different as local exchange carriers hold considerable market power over the provision of the former while the latter is subject to considerably more competitive pressure.

[47] See Taylor (1994).

No regulation: Obviously, where a firm's prices are set (and changed) by regulatory fiat, any empirical examination of pricing behavior must take into account the regulatory authority's actions, as well as the firm's behavior in setting output prices generally. As Peltzman (1976) and subsequent authors have explained, there is a pronounced tendency for regulators to spread both costs and benefits, emanating from whatever sources, across constituents in relation to the political influence wielded by the individual interest groups involved – both firms and customers. As a result, the cost reduction caused by an input price reduction is unlikely to be fully passed through to consumers under regulation. In addition, a given cost reduction is unlikely to be passed through to all consumer groups, e.g., business and residential, equally. Therefore, where regulation is in force, as it has been throughout most of the post-divestiture period, the pass through test is likely to fail to yield valid information regarding the intensity of competition[48].

Also of importance, even where regulation has been relaxed or even removed, observed prices of individual products are likely to continue to reflect prior regulatory pricing decisions, e.g., cross-subsidization policies, for some time. Suppose, for example, that, under regulation, product A was subsidized in the Faulhaber (1975) sense, while product B was providing the source of that subsidy. Following deregulation and assuming competition, A's price should rise and B's price should fall. Neither price, however, is likely to move to the competitive level instantaneously[49]. As a result, as these prices adjust to their equilibrium values, they are unlikely to exhibit identical, or even similar, responses to access charge reductions. Thus, prior regulatory controls are likely to distort subsequent pass through behavior.

Other complications: The above list does not exhaust all of the difficulties encountered in attempting to draw inferences about the competitiveness of a market from observed pass through performance. For example, it is obvious that any study of pass through behavior must control for other contemporaneous sources of cost shifts, e.g., other input price changes or technological advances. In addition, Beard (1999) lists several other complications that can plague this approach.

[48] As noted earlier, over the entire post-divestiture period, varying degrees of regulation of AT&T and other carriers have been imposed at both the state and federal levels. During this period, state public service commissions and the FCC have substantially altered their regulatory policies – in some cases, several times. Rate-of-return regulation, rate bands, price caps, sharing formulas, and deregulation have all been employed at various times in various jurisdictions. Therefore, regulation and changes in regulation provide an additional source of ambiguity in trying to make inferences about the status of competition from observed pass through behavior.

[49] This is particularly true, of course, where regulatory controls are removed incrementally over time, as has been the case in the telecommunications industry. In addition, even after "deregulation," regulatory officials may continue to monitor firms' pricing performance and take actions to influence that performance.

In light of the foregoing considerations, the pass through methodology does not appear to provide a reliable technique for accurately measuring the intensity of competition in most real world markets and, particularly, in the long-distance market. Proposition 1, above, does not appear to hold generally.

4.2. Empirical analyses

Turning to Proposition 2, there has been an ongoing debate by interested parties over the past several years regarding the observed pass through performance of long-distance firms. These parties have attempted to either support or refute the hypothesis that AT&T and other interexchange carriers have failed to fully pass through the access charge reductions they have experienced. At first blush, one might expect that this type of empirical question, irrelevant though it may be, would be easily answered by a straightforward analysis of the relevant data. Such an expectation, however, has not been met. Arguments and counterarguments have flown back and forth on this issue for several years.

In addition to the unavoidable biases created by the adversarial process, there are several reasons for the inability to answer this question more definitively. First, given the complexities of the tariff structures for both output prices and access charges, as well as the plethora of discount programs for the former, accurate and uncontroversial measurement of changes in either is not an entirely straightforward exercise. For example, should the observed price change be calculated as an average across all customers, including both residential and business, or should it focus on more narrowly defined customer groups? Should it relate to basic tariffed rates or to the companies' average revenues? These and other questions tend to complicate the analysis and appear to have led to conflicting results.

Second, in addition to these measurement problems, there are difficulties involved in attempting to hold other influences on costs and prices constant. For example, other, non-access input prices can change, either reinforcing or offsetting the impact of a change in access charges on costs. Also, technological change has been occurring quite rapidly in this industry, further complicating the relationship between access charges and costs.

Third, there is also an issue of timing. Specifically, over what interval of time should the relevant price and cost changes be measured? Even in competitive markets, prices are not likely to adjust instantaneously to a shift in costs. Therefore, some lags may occur that have nothing to do with the presence of market power. Moreover, due to regulatory announcements, firms often know that access charges are going to be lowered before the fact. In an attempt to gain an edge over competitors, these firms may reduce their prices before the actual reduction in access charges takes effect. In that event, a researcher examining observed price changes after access charge reductions becomes effective might erroneously conclude that no adjustment has occurred.

Despite these empirical complexities and the serious theoretical infirmities associated with this approach, several studies have appeared which attempt to draw inferences regarding the intensity of competition in long-distance markets from observed access charge pass through behavior. The first of these was Taylor and Taylor (1993). Relying upon observed changes in AT&T's annual revenues from its interstate tariffed switched services and its annual access charge payments, these authors find that real revenues for interstate services fell at a rate of approximately 2.2 percent per year in the 1984–1992 period after adjusting for access charge reductions[50]. Standing alone, this finding suggests that AT&T's aggregate real output price reductions have exceeded its access charge reductions during the post-divestiture period. At the same time, however, these authors also find that access-adjusted real toll rates had fallen at a faster pace, about twice the rate, during the decade preceding divestiture. From this evidence, Taylor and Taylor conclude that:

> "These findings are hardly consistent with the view that competition among interexchange carriers led to drastically lower prices.... Since reductions in carrier access charges represent reductions in marginal costs for all long-distance companies, these data are more consistent with the presence of a regulated price umbrella than a competitive market[51]".

Thus, this study views a slowdown in the rate at which pass through occurs as evidence of non-competitive market performance. In so doing, Taylor and Taylor rely upon what may be referred to as a 'second derivative' argument. Specifically, they ignore the fact that the *level* of prices for long-distance services have declined, consistent with competition. Similarly, they ignore the fact that prices declined in response to cost changes (the first derivative), consistent with competition. Instead, curiously, they base the inference of a lack of competition on the observation that the rate of change of price changes in response to cost changes (the second derivative) declined following the divestiture. We know of no theoretical basis for such a view.

In a second related paper, Taylor and Zona (1997) present a somewhat more broad-based analysis of the extent of competition in the interstate long-distance market. Here, seven different indices of market performance are proposed and evaluated in an attempt to draw inferences about the competitiveness of that market. The most extensively analyzed of these indices, however, is AT&T's observed access charge pass through behavior. We focus here on that portion of the paper.

Taylor and Zona examine AT&T's pass-through performance over the period from divestiture through the first quarter of 1995 (prior to the FCC's decision

[50] As these authors acknowledge (footnote 5, p. 187), this focus on AT&T's tariffed services will tend to understate reductions in actual transaction prices, because such rates fail to reflect certain bulk business service offerings and residential discount calling plans. Moreover, this bias is likely to have increased over time as these discounted services have increased both in number and subscribership.

[51] Taylor and Taylor (1993, pp. 186–187).

removing price cap regulation of AT&T in October of 1995). That examination consists of three separate, but related approaches. First, as in the earlier paper by Taylor and Taylor (1993), these authors calculate nominal revenue and access payment changes over the post-divestiture period (updated to Q1: 1995).

Second, they calculate Laspeyres price indices for access charges and a subset of AT&T's toll prices over the same period. Third, they calculate changes in AT&T's average revenue per minute (ARPM) and access charges. This last approach is applied to both AT&T's overall ARPM over the decade following divestiture and AT&T's ARPM for its so-called Basket 1 services over the 1989–1994 (price cap) period[52].

From these various calculations, Taylor and Zona find that: (1) overall, after adjusting for inflation, AT&T has fully (or more than fully) passed through its access charge reductions, although the percentage 'excess' pass through has been lower than that observed in the pre-divestiture (1972–1983) period; and (2) in the post-1989 (price cap) period, AT&T has not fully passed through all of its access charge reductions in the prices charged for its Basket 1 services. From these findings, these authors conclude that:

> "Based upon our investigation, we do *not* find evidence of effective competition in the interstate long-distance market; rather, we find evidence more consistent with noncompetitive behavior" (p. 229).

A third study addressing the pass through issue is contained in Crandall and Waverman (1995, Chapter 5). This study differs from the approaches employed by Taylor and Taylor (1993) and Taylor and Zona (1997) in at least two respects. First, the Crandall and Waverman analysis focuses upon AT&T's tariffed (i.e., undiscounted) intrastate interLATA toll rates and intrastate access charges over the 1987–1993 period. This sample contains data from twenty-nine to thirty-three multi-LATA states, depending upon availability of the relevant price data. Second, this study attempts to control for the effects of varying regulatory constraints (over both states and time) by specifying an econometric flow through model that: (1) includes a set of binary variables for the types of regulation applied in these various jurisdictions at various points in time; and (2) is estimated separately across four different mileage bands. The first feature of this modeling approach allows observed pass through behavior to be influenced by the overall stringency of the existing regulatory constraint. The second feature allows pass through behavior to vary across mileage bands. Such disaggregation permits certain prior regulatory pricing distortions, such as the cross-subsidization from long-haul to short-haul toll calls, to affect observed pass through performance.

[52] Under the FCC's price cap plan, Basket 1 services consisted primarily of relatively low volume residential and small business MTS services.

Several interesting findings flow from the empirical results presented in this study. For example, the results suggest that regulation affects flow through behavior and the growth of competition in several ways. First, the nature or stringency of the regulatory regimes adopted by the various states in the sample—specifically, the degree of pricing flexibility afforded AT&T—is shown to influence the extent to which access charge reductions are passed through to final output prices. Specifically, Crandall and Waverman (1995, p. 161) state that:

> "In general, regulatory flexibility increases the degree of pass-through of access cost, whereas no flexibility reduces it."

Thus, any structural analysis of pass through behavior should attempt to control for regulatory influences.

Second, current flow through behavior is shown to be influenced by prior regulatory-induced pricing distortions. This conclusion stems from the divergence in the empirical results across the different mileage bands[53]. Specifically, substantially greater flow through of access charge reductions is observed for the longer-distance calls. Because it is generally acknowledged that short-haul toll rates tended to be relatively under-priced under regulation, movement toward more competitive pricing would necessitate this result—a relative increase in the prices observed in the lower-mileage bands[54].

As noted earlier, where regulation is accompanied by cross-subsidization, as it almost always is, deregulation and the growth of competition will have markedly different effects on individual service prices, or, in this case, individual components of a given service's price structure. As a result, valid inferences about the status of competition cannot be drawn from observations of access charge pass through performance with respect to an individual service (or subset of services).

Third, as previous authors have concluded, Crandall and Waverman's results generally support the argument that, where competition has become viable, a continuation of regulation tends to retard the growth of competition[55]. To the extent this effect occurs, a Catch-22 emerges. Policy makers are reluctant to deregulate a market because they are unconvinced that competition is sufficiently vigorous to protect consumers from monopolistic pricing. Yet, continued regulation prevents the market from demonstrating the sort of competitive performance that would allay these fears. The result is a continuation of regulation,

[53] Crandall and Waverman (1995, p. 159) write that: "Disaggregating results across mileage bands is clearly important."

[54] At pages 160–161, these authors write: "The results for the shortest mileage bands are clearly the least precise. This may be because rates in these bands were below competitive-equilibrium rates before deregulation by the states. (Rates net of access charges in the shortest band were near zero in 1987)."

[55] At page 163, these authors state that: "It is also clear that the continuation of regulation impedes competition."

perhaps well beyond the point at which net social welfare considerations would warrant deregulation[56]. We suspect that this phenomenon goes a long way toward explaining the slow pace of deregulation of long-distance markets at both the state and federal levels.

Finally, from the pass through evidence they examine, Crandall and Waverman conclude that the intrastate interLATA long-distance market is reasonably, but not perfectly, competitive. Specifically, these authors write that:

> "The combination of all of these results suggests to us that considerable competition exists in the interstate [we believe they mean intrastate] market, but not perfect competition"[57].

This finding is clearly at odds with the conclusions drawn in the earlier studies by Taylor and Taylor (1993) and Taylor and Zona (1997).

Another study that addresses the pass through issue, although somewhat tangentially, is by Edelman (1997). This study examines AT&T's basic schedule, undiscounted, residential rates over the 1980–1992 period for fifteen sample calls specified by time-of-day, duration, and distance. Given these rates, Edelman imputes AT&T's total annual billings for the fifteen sampled calls and, from 1984, the total access charges that AT&T paid to local exchange companies for completion of these calls. AT&T's billings, net of access charges, can then be calculated for each year from 1984–1992.

From these calculations, Edelman (1997) finds that AT&T's basic schedule residential rates declined by almost eleven percent, net of access charges, over the sample period. Thus, over the entire period of observation, AT&T's basic schedule rate reductions exceeded its access charge reductions. Edelman also finds, however, that during the first three years that AT&T was subjected to price cap regulation (1989–1992), its rates, net of access charges, increased. Of course, this finding may be due to the study's focus on undiscounted rates only, as this is the period during which subscription to discount programs was increasing rapidly.

Finally, the most recent empirical study to appear on this issue is one conducted by Hill and Beard (1999) on behalf of MCI WorldCom. These authors make use of a publicly available data source providing residential customers' interstate long-distance bills. These data cover the 18-month period from January 1997, through July 1998, a period during which the FCC substantially lowered interstate access charges on two separate occasions.

Hill and Beard employ a bootstrap analysis of these data in order to allow, for the first time, formal statistical tests, not dependent upon specification of a structural model, to be conducted of the extent of flow through. They define flow-through as follows:

[56] There are, of course, other potential explanations of why regulatory policies generally tend to outlive their usefulness. See, e.g., McCormick, Robert E., William F. Shughart II, and Robert D. Tollison (1984).

[57] Crandall and Waverman, op.cit., p. 163. The statement within the parentheses was added by us.

> "Thus, flow-through occurs, assuming other costs constant, if the change in ARPM equals the change in average access charges per minute. This ensures that no 'substantial new profits' accrue to the long distance industry." [Emphasis in original.]

Importantly, this definition is consistent with Beard's (1999) earlier theoretical analysis which emphasized the impact on sellers' overall profits rather than individual buyers' prices in a multiproduct setting[58]. At the same time, these authors recognize the thin theoretical ice upon which pass through studies must skate, pointing out that:

> "Our statistical analysis presumes that measuring flow through on a period-to-period basis has merit. In fact, there are good reasons to believe that such an analysis is inappropriate."

Nonetheless, given this caveat, Hill and Beard's statistical analysis reveals that:

> "In every case, we find results statistically consistent with 100 percent pass-through or better for either estimate of access charge reductions... We conclude, with 90 percent certainty that MCI customers enjoyed reductions in costs per minute equal to, or exceeding, access charge reductions"[59].

Moreover, this same analysis has recently been replicated by these authors for AT&T yielding equivalent results. Thus, whatever the theoretical limitations of this sort of analysis, it appears that, at least in the most recent period studied, at an aggregate level, using average revenue to measure price, complete, or more than complete, pass through of access charge reductions has occurred. Proposition 2, above, then appears to be incorrect as well.

4.3. Summary

Overall, in our opinion, the access charge pass through debate has not provided a great deal of insight regarding competitiveness of the interLATA long- distance market. Nonetheless, a considerable amount has been written on this subject, and it has unquestionably influenced a number of regulatory decisions pertaining to both access charge reductions and RBOC entry into the interLATA market. Thus, despite its weak theoretical underpinnings and the questionable nature of some of the empirical analyses, this topic warrants our attention here.

While the empirical evidence strongly suggests that this market passes the pass through test, the theoretical significance of that finding is not altogether clear. There are simply too many ambiguities involved in trying to connect observed

[58] On page 4 of their report, Hill and Beard (1999) write: "This definition of flow-through is required if flow-through is to be used as a rough proxy for the degree of competition in the long-distance industry. In competitive markets, economic profits tend to zero, but there is no general rule that says all prices charged by a multi-product firm will exactly reflect a change in unit cost, whatever the degree of competition... Thus, economic theory implies that no simple pass-through relationship for a given tariff (among many) will exist."

[59] Hill and Beard (1999, pp. 22–23).

pass through performance to the competitiveness of a market to place a great deal of weight on these results. What is clear, however, is that this market's pass through performance does not warrant an inference of any significant exercise of market power on the part of the interexchange carriers. Whether it unambiguously demonstrates intense competition, however, is another question.

5. NEIO studies

A fourth approach that has been used to assess the intensity of competition in the long-distance market applies empirical techniques which Bresnahan (1989) classifies collectively as the New Empirical Industrial Organization (or NEIO) approach[60]. These techniques generally focus on either individual firm or industry time series data – as opposed to the more traditional empirical methods in the Industrial Organization literature, which have primarily employed inter-industry cross-sectional data. These less aggregate time series data are then used to estimate structural equations that are derived from explicit theoretical models of the industry being examined.

Under certain conditions, market events can produce data that allow inferences to be drawn regarding the percentage departure of price from marginal cost, even though the relevant costs are not, themselves, generally observable. When such events occur, fairly generalized models of industry demand functions and individual firms' supply relations can yield estimates of structural parameters that shed light on the type of behavior exhibited by market participants. Specifically, such parameter estimates can indicate: (1) the price elasticity of an individual firm's residual demand function; and/or (2) the degree of coordination between firms in the market, i.e., the conjectural variations exhibited by these firms.

Thus, properly conducted, NEIO studies are able to assess either unilateral or collective market power resulting from either single-firm dominance or multi-firm coordination. Both types of studies have been conducted on the long-distance industry. The following subsections report the results of each.

5.1. Residual demand studies

Two separate NEIO-type studies have attempted to estimate the price elasticity of AT&T's residual demand curve in the post-divestiture period. The first of these is by Simran Kahai and the authors of this chapter[61]. This study employs a dominant firm/competitive fringe model to characterize the long-distance industry over the 1984:3 through 1993:4 period, arguing that the assumptions of that model were reasonably well satisfied in this market during that period. Assuming that to

[60] See Bresnahan (1989) and the studies cited therein.
[61] Kahai, Kaserman, and Mayo (1996).

be the case and viewing AT&T as the dominant firm, it is well known that the price elasticity of this firm's demand will be given by:

$$\eta_{ATT} = (\eta_M / S_{ATT}) + [(1 - S_{ATT})/S_{ATT}]\varepsilon_F \tag{11}$$

where η_{ATT} is AT&T's residual demand elasticity, η_M is the market demand elasticity, S_{ATT} is AT&T's market share, and ε_F is the price elasticity of fringe firms' collective supply[62].

Equation (11) implies that the intensity of competition faced by AT&T is determined by the above three parameters – η_M, S_{ATT}, and ε_F. Prior estimates of η_M are available in the published literature, and AT&T market share figures (based upon both minutes-of-use and transmission capacity) are available as well. What is lacking, then, is an estimate of ε_F. To obtain that estimate, Kahai et al. specify a two-equation simultaneous system of fringe supply and market demand. Such a system not only allows estimation of the missing supply elasticity, ε_F, but also provides an additional empirical benchmark for the elasticity of total market demand, η_M.

Estimation of this model with two-stage least squares using quarterly (interstate residential) observations over the 1984:3 through 1993:4 period (38 observations) yields an estimated fringe supply elasticity at the sample means of 4.38 and a market demand elasticity of -0.49[63]. Using these estimates along with two prior estimates of AT&T's market share – 60 percent on the basis of output (minutes-of-use) and 40 percent on the basis of assets—yields residual demand elasticities for AT&T of -3.73 and -7.81, respectively. These elasticities, in turn, yield Lerner index values of 0.29 and 0.13, which are at the lower end of the theoretical range of values for this index[64].

Finally, Kahai, Kaserman, and Mayo compare these Lerner index estimates to prior estimates [obtained from Hall (1988) and Bresnahan's (1989) survey] for other industries in the U.S. On the basis of this comparison, the authors conclude that:

> "... relative to other firms in the U.S. economy, AT&T possesses very little market power"[65].

The second NEIO-type study to estimate AT&T's (and, in this case, MCI and Sprint's combined) residual demand elasticity is by Ward (1999). Like Taylor and Taylor (1993), Ward finds that the available data are not adequate to estimate the desired residual demand elasticities directly. Consequently, he focuses upon estimation of firm-level Marshallian demand functions, which reveal the relationship

[62] See Saving (1970), and Landes and Posner (1981).

[63] The latter figure falls at the low end of the range of interLATA toll demand elasticity estimates reported by Taylor (1994).

[64] Kahai et al. (1996, pp. 513–514) discuss several caveats associated with these estimates, most of which tend to bias the calculated Lerner index values upward.

[65] Kahai et al, p. 513.

between the firm's quantity demanded and price, assuming all other firms hold their prices constant.

Given an estimate of the firm's Marshallian own-price elasticity, $\eta_{ii,}$ and the cross-price elasticity with respect to other firms, η_{ij}, residual demand elasticities, η_{ii}^{R}, are calculable from

$$\eta_{ii}^{R} = \eta_{ii} + \Sigma \eta_{ij} \theta_{ji}^{R}, \tag{12}$$

where $\theta_{ji}^{R} = (\partial P_j / \partial P_i)(P_i / P_j)$ is firm j's price reaction to firm i's price changes in percentage terms (i.e., the elasticity of firm j's price with respect to firm i's price). Thus, Ward is able to use Equation (12) to obtain estimates of η_{ii}^{R} from his econometric estimates of η_{ii} and η_{ij} and various plausible assumptions regarding the value of θ_{ji}^{R}. Because $\eta_{ij} > 0$ (competing firms' outputs are substitutes), where firms are able to coordinate their pricing decisions, $\theta_{ji}^{R} > 0$ as well, and residual demand will be less elastic (lower in absolute value) than Marshallian demand.

Ward uses a two-level budgeting approach to estimate the firm-level Marshallian demands. Under this approach, consumers are assumed to first decide how much long- distance service to purchase, and then decide which long-distance firm to purchase that service from. The former, upper-level long-distance demand function is estimated from monthly interstate data over the July 1986, through August 1991, period. The lower-level individual firm demands are then estimated with both interstate and a five state sample of intrastate monthly data. Time periods covered are from January, 1988, through December, 1991, for the interstate data, and January, 1988, through October, 1991, for the intrastate data.

Aggregate market quantities and revenues are obtained from billing records of the local exchange carrier serving the five-state region, Southwestern Bell, for both switched access minutes and special access lines. Dividing revenues by outputs, minutes for switched access and lines for special access, yields average aggregate prices for these access services. Individual firm-level output prices are constructed from the tariff filings of AT&T, MCI, and Sprint with the FCC and the relevant state public service commissions. Finally, Ward estimates both an immediate adjustment and a two-year partial adjustment model with these data.

At the industry level, these estimates yield a long-run price elasticity of demand of -0.89, which is at the high end, but not outside the range, of prior estimates. At the firm level, lower-bound Marshallian own-price elasticity estimates range from -5.34 to -10.59 for AT&T and from -15.22 to -26.83 for MCI and Sprint, depending upon which data set, intrastate or interstate, is used and whether an immediate adjustment or a partial adjustment polynomial distributed lag structure is employed. From these results, Ward (1999, p. 667) concludes that:

> "...consumer behavior appears to reflect the belief that service from AT&T, MCI, and Sprint were close substitutes."

Given these estimates, Ward then examines the likelihood of coordinated pricing behavior. He begins by arguing that "some characteristics of the long-distance industry suggest that AT&T's rivals are not likely to match an AT&T price change"[66]. He then explains several features of this market that would appear to preclude, or, at least, reduce the likelihood of, highly coordinated pricing behavior.

Following this discussion, Ward calculates the potential deadweight loss due to supracompetitive pricing by AT&T as a percent of industry revenue under a wide range of assumptions regarding the value of the price reaction term in equation (12), above. These calculations lead him to conclude that:

> "For plausible assumptions regarding the degree of price coordination among firms, AT&T's residual demand elasticity is likely to be below [i.e., more elastic than] -5"[67].

Thus, Ward's residual demand elasticity estimates for AT&T are quite consistent with the earlier estimates provided by Kahai, et al (1996). Both sets of estimates suggest the presence of reasonably effective competition.

5.2. *Conjectural variation studies*

A firm's residual, or perceived, demand takes into account both the reaction of consumers in general and the reaction of rival producers when the firm alters either its price or output. That is, residual demand elasticity reflects both the total market demand elasticity and the firm's expectations regarding rival firms' responses[68]. The latter term – the conjectural variation parameter – indicates the degree of interfirm price, or output, coordination present in observed market outcomes. As Bresnahan (1989) explains, it is possible to estimate this parameter from existing market data under certain conditions. The results of that estimation, then, can provide insight regarding the extent to which rivals are able to act in concert to raise the market price above marginal costs, i.e., to exercise collective market power.

At least one paper, Taylor and Zona (1997), has attempted to estimate the conjectural variation parameter implicit in AT&T's observed pricing behavior. Following Bresnahan (1989), these authors begin with the first-order condition for individual firm profit maximization which they write as:

$$P_t = c_t/(1 + \theta/\varepsilon_t) \tag{13}$$

where P_t is price at time t, c_t is (constant) marginal cost, ε_t is the market demand elasticity, and θ is the conjectural variation parameter. This last term, θ, is the primary object of estimation.

[66] Ward, op.cit., p. 673.
[67] Ibid., p. 567.
[68] See Bresnahan (1989) and equation (12), above.

To carry out that estimation, Taylor and Zona first rewrite (13) as[69]:

$$\ln(p_t) = \ln(c_t) + \ln(1 + \theta/\varepsilon_t) \tag{14}$$

They then attempt to proxy c_t with two separate variables, an index of access prices and a simple time trend. The latter variable is intended to measure "... other components of cost that trend smoothly over time..."[70] No justification is provided for the implicit assumptions that: (1) the time trend variable is highly correlated with non-access cost changes; and (2) that variable is not correlated with other, excluded variables.

Also, perplexingly, these authors indicate that "ε_t is the price elasticity of demand for a good at time t calculated using Ward's results"[71]. This use of a prior econometric estimate of the price elasticity of demand for AT&T's service as a right-hand variable in Equation (14) begs the question of how this prior elasticity estimate can vary over the authors' sample period. In addition, this use of an econometric estimate as an explanatory variable introduces an errors-in-variables problem that is likely to bias the resulting parameter estimates.

Fitting this revised version of Equation (14), Taylor and Zona obtain an estimate of θ equal to 2.55, which is significant at the 0.01 level[72]. From this estimate, the authors conclude that:

> "... our results reject the hypothesis that AT&T acts as a competitive firm in the interstate long-distance market. Thus, if its prices are not constrained by regulation or the threat of antitrust intervention, then AT&T would wield substantial market power because of coordinated pricing behavior among the IXCs[73].

This conclusion, then, tends to support MacAvoy's (1995) contention, discussed earlier, that the major long-distance carriers – AT&T, MCI, and Sprint – have successfully coordinated their pricing behavior in order to raise and maintain market price above marginal cost. That is, both of these papers argue that tacit collusion is present in this industry.

That conclusion, however, is inconsistent with much of the prior literature in this area[74]. In order to investigate this issue further, we have developed and

[69] The plus sign is apparently a typo, as a minus obviously should precede the last term.

[70] Taylor and Zona (1997, p. 243).

[71] Taylor and Zona (1997, p. 243) indicate that the source for these elasticity estimates is the study by Michael Ward reported in Reply Comments of the Staff of the Bureau of Economics of the Federal Trade Commission, In the Matter of Revisions to Price Cap Rules for AT&T, CC Docket No. 93-197, October 25, 1993.

[72] The data used for this estimation are Laspeyres indices for output and access prices constructed by the authors. These indices are calculated quarterly from 1984:3 through 1994:4.

[73] Taylor and Zona (1997, p. 244).

[74] See our earlier discussion of this issue in Section 2, above. In addition, the results reported in Section 3 regarding the impact on price of reduced regulation also tend to reject the tacit collusion hypothesis.

applied an alternative approach to detect the emergence of cooperative behavior in this market. That approach is presented next.

5.3. An alternative test for tacit collusion

As noted earlier, MacAvoy (1995) argues that the nature of behavior exhibited by the three major long-distance carriers – AT&T, MCI, and Sprint – shifted dramatically around the 1989–1990 time frame. Specifically, he argues that the FCC's adoption of price cap regulation of AT&T, along with the implementation of equal access, created market conditions that were much more favorable to coordinated pricing behavior by these three firms.

Under this theory, then, MCI and Sprint were encouraged both by market events and regulatory change to alter their behavior from rivalrous to accommodating. Both the lure of profits and the, now-credible, threat of retaliation by AT&T is alleged to have motivated this change. The resulting shift in these firms' reactions to AT&T's pricing decisions is described by MacAvoy (1995):

> "When one firm was making almost all sales and regulatory conditions required a floor under that firm's prices, then the best pricing strategy of the other two firms most plausibly was that designated to increase their market shares... But when shares of the second and third largest firms increased to levels comparable to that of the first firm, and price floor regulation was eliminated, the second and third firms would not take further individual initiatives to increase their shares."

If, in fact, the fundamental nature of MCI and Sprint's reaction functions changed in 1989, then some indication of that change should be discernible in the market data which reflect these companies' price/output, i.e., supply, decisions. In microeconomic terms, the sort of structural shift described by MacAvoy (1995) would correspond to an alteration of the market equilibrium from a simple dominant firm/competitive fringe model to a (partial) conspiracy model in which some formerly fringe producers begin to coordinate their price and output decisions with those of the dominant firm. Rather than supplying output along their individual marginal cost curves, these firms should begin to restrict output in concert with the dominant firm in order to raise price above marginal costs.

To test for this sort of structural shift in the supply response of these two firms, which collectively produce a large portion of the total fringe, i.e., non-AT&T, output, we extend our earlier analysis (Kahai, Kaserman, and Mayo [1996]) of the competitive fringe supply curve for residential interstate toll service[75]. In our previous study, we estimated simultaneously the fringe supply and market

[75] Proponents of the tacit collusion argument generally acknowledge that the intensity of competition is greater in the business services segment of the industry. Thus, if non-competitive performance is present, it should be most apparent in the residential service segment.

demand for residential interstate service. In that study, the inverse fringe supply curve was specified as

$$P = P_F(Q_F,\ PA,\ EA), \tag{15}$$

where Q_F is the output of the competitive fringe (measured as the number of minutes sold), PA is the per-minute price of carrier access, and EA is the percent of telephone lines converted to equal access. Theoretical considerations described in our prior paper suggest that $\partial P_F/\partial PA > 0$ and $\partial P_F/\partial EA < 0$ in this supply equation.

The expected sign of $\partial P_F/\partial Q_F$, however, is contingent upon the type of behavior exhibited by fringe producers. Specifically, in an industry where tacit collusion occurs, a price increase by the dominant firm, which, for a given demand, requires a reduction in that firm's output, will be met by a corresponding, though not necessarily proportional, reduction in the output of fringe producers. This sort of accommodating supply response under tacit collusion is emphasized by Posner (1976):

> [in tacit collusion,] "one seller communicates his 'offer' by reducing output, and the offer is 'accepted' by the actions of his rivals in restricting output as well[76]".

Conversely, in the absence of tacit collusion, a reduction in the dominant firm's output, or an increase in the dominant firm's price, will be met by an increase in the output of the fringe. Such divergent supply responses of the competitive fringe under these two alternative modes of conduct should allow us to distinguish empirically a shift from competitive (non-cooperative) behavior to collusive (cooperative) behavior. Therefore, the appearance of tacit collusion can be detected in our model by a structural shift in the competitive fringe supply curve at the time such collusive behavior is believed to have materialized.

Accordingly, we respecify the inverse fringe supply equation here to allow for the possibility of a shift in the responsiveness, i.e., elasticity, of the fringe firms' supply beginning in 1989, the time at which the FCC implemented price-cap regulation of AT&T – the event that Professor MacAvoy alleges created conditions favorable to tacit collusion in the industry. Such a change may appear as a shift in either the slope or intercept of the $P_F(.)$ function in Equation (15). Writing this function in linear form, we have:

$$P = \beta_0 + \beta_1 Q_F + \beta_2 PA + \beta_3 EA. \tag{16}$$

Then, defining C to be a binary variable that becomes one in the third quarter of 1989 and is zero prior to that time, we can allow both the intercept and slope of the fringe supply curve to shift by specifying

$$\beta_0 = \delta_0 + \delta_1 C, \tag{17}$$

[76] Posner, (1976, p. 72).

and

$$\beta_1 = \alpha_0 + \alpha_1 C. \tag{18}$$

Substituting Equations (17) and (18) into Equation (16), we have:

$$P = \delta_0 + \delta_1 C + \alpha_0 Q_F + \alpha_1 C \cdot Q_F + \beta_2 PA + \beta_3 E. \tag{19}$$

If tacit collusion emerged in this industry in 1989, then the slope shift parameter α_1 in Equation (19) should be positive and significant[77]. That is, the fringe supply curve should become less price elastic – signifying a less competitive response by these firms to a price change – following an agreement, tacit or otherwise, to behave in a more cooperative fashion. Thus, a straightforward test of the tacit collusion hypothesis can be carried out by estimating Equation (19) and calculating the relevant (pre- and post-1989) fringe supply elasticities. If tacit collusion emerged in the latter period, fringe supply elasticity should fall significantly.

The data utilized to conduct this estimation are quarterly observations for residential interstate calling over the 1984: 3 through 1995: 4 period compiled by the FCC[78]. They are the same data used in our earlier 1996 paper, updated with more recent observations.

Table 2 provides variable definitions, descriptive statistics and data sources. Given a DF/CF model, or any other model of imperfect competition, fringe output and price are simultaneously determined. Thus, Equation (19) was estimated with two-stage least squares (2SLS), with the market demand variables employed in our prior paper used for identification purposes. There is no need to estimate that demand equation here, because our focus is on the fringe supply elasticity.

Our estimation results are reported in Table 3. The overall explanatory power of the model is quite high, and all coefficients for which a clear hypothesis exists attain statistical significance and the anticipated signs. Specifically, consistent with our previous findings, we obtain a positively sloped inverse fringe supply curve that shifts upward, a reduction in supply, with increases in the price of carrier access and downward, an increase in supply, with increases in the percentage of lines converted to equal access.

The result that is of primary interest here is the coefficient estimate associated with the interaction variable, $C \cdot Q_F$. If, as has been alleged, tacit collusion emerged in this industry in 1989, this coefficient should be positive and significant. Instead,

[77] The sign of the intercept shift parameter δ_1, however, is theoretically indeterminate. We incorporate this parameter in our specification in order to place as few constraints on the model as possible. The results we obtain are not sensitive to whether this additional parameter is included.

[78] We end the sample period at this latter point because, after 1995:4, price cap regulation of AT&T was eliminated. This truncation prevents the data from contamination due to another shift in regulatory policy.

Table 2
Descriptive statistics

Variable	Definition	Mean	Var	Minimum	Maximum	Source
C	Indicator variable with values equal to one beginning the third quarter of 1989, 0 otherwise.	0.600	0.245	0	1	(a)
Q_F	Interstate switched-access minutes by carriers other than AT&T	27.265	183.563	5.900	53	(b)
PA	Real price per minute of total access charges per conversation minute	0.096	0.001	0.06	0.177	(c)
P	Average daytime real price per minute of AT&T's long-distance interstate telephone service for a 10 minute, 200 mile call	0.263	0.003	0.21	0.409	(c)
EA	Percentage of total industry lines converted to equal access.	76.944	734.544	0	98.900	(b)

Sources: (a) authors' construction; (b) Federal Communications Commission, Statistics of Common Carriers, 1995, 1996 Edition; (c) Federal Communications Commission, Common Carrier Bureau, Industry Analysis Division, Reference Book: Rates, Indexes, and Household Expenditures for Telephone Service, May 1996.

we find that it is negative and significant at the 0.05 level, suggesting a reduction in the slope of the inverse fringe supply curve. At the same time, however, the intercept shift coefficient is positive and significant at the 0.01 level, indicating a reduction in fringe supply.

Table 3
Inverse fringe supply equation: 2SLS estimates

Variable	Coefficient	*t*-Statistic
Intercept	0.035	1.030
C	0.034	1.860***
Q_F	0.003	3.890*
$C \cdot Q_F$	−0.002	−2.170**
PA	1.954	10.740*
EA	−0.001	−5.770*

$R^2 = 0.9947$
* Significant at the 0.01 level
** Significant at the 0.05 level
*** Significant at the 0.10 level.

Together, however, these structural shifts increase overall fringe supply elasticity, thereby indicating an increase in the intensity of competition exerted by the fringe in this market. Calculating fringe supply elasticity at the point where the pre-1989 and post-1989 supply curves intersect, in order to calculate these elasticities at a common price and output, we find that, prior to 1989:2, this elasticity was 2.78; and, after 1989:2, it is 5.28[79]. Clearly, such a structural shift in the elasticity of fringe supply is inconsistent with, and indeed it contradicts, the hypothesis that tacit collusion emerged in this industry at this time. On the contrary, it suggests an increase in the intensity of competition in the latter period.

Overall, the NEIO studies surveyed here, along with our own empirical investigation of the tacit collusion issue, tend to suggest that the long-distance market is subject to reasonably intense competition. While one can find evidence that this market does not conform exactly to the dictates of the model of perfect competition, it nonetheless appears to exhibit a degree of competition that is equal to or greater than that shown by numerous other unregulated markets.

6. Competition for international calling

To this point, the studies we have reviewed have focused exclusively on domestic long-distance calling – i.e., interexchange calls that both originate and terminate within U.S. boundaries. While the domestic segment of the overall long-distance market is, by far, the more important source of U.S. long-distance companies' revenues, accounting for just over 80 percent of total toll revenues in 1997, the international segment has been growing rapidly in relative importance. This segment's share of revenues has increased from only 1 percent in 1950 to over 20 percent in 1997, with most of that increase occurring since the mid-1980s[80]. In addition, this segment of the market has begun to receive some limited attention in the literature assessing competitive intensity. Accordingly, we will briefly review that emerging literature here.

Before turning to the individual studies, it is worth pointing out that, in all likelihood, the growth of competition in the international segment of the U.S. market has lagged the growth of competition in the domestic segment. There are a number of possible explanations for this slower movement to competition. Probably most important, incumbency advantages inherited by AT&T at divestiture appear to have been more substantial in this segment.

In particular, the ability of new entrants in the U.S. market to provide international services is contingent upon their ability to negotiate an operating agreement with correspondent carriers in the countries to which they wish to

[79] These elasticities were calculated at the sample mean values of the exogenous variables PA and EA.
[80] All figures reported here are from Blake and Lande (1999), which provides a very useful overview of this segment of the market.

terminate calls. Separate agreements are required for each country. In many cases, these foreign carriers have been government-owned monopolies with little interest in promoting competition in the U.S. market. As a result, AT&T's share of the international segment of the market failed to decline as rapidly as its share of the domestic segment. As late as 1990, AT&T still maintained an international market share of 70 percent. By 1998, however, AT&T's share of international calling revenues from facilities-based carriers and facilities resellers had declined to roughly 53 percent.

In addition, transmission capacity costs have been quite large for trans-oceanic calls, although these costs have been falling in recent years. As a result, much of the new capacity that has been put in place in this segment has involved joint ventures among the larger U.S. carriers, creating potential concerns regarding the intensity of competition among the participants.

For these and other reasons, there appears to be a general concensus that new entrants into the long-distance market in the U.S. have faced somewhat greater obstacles in penetrating the international segment. Consequently, the intensity of competition found in this market segment is likely to provide a lower-bound estimate of the competitiveness of the overall long-distance market. With this observation in mind, we turn to the literature on this topic.

Despite the overall size of the international segment of the U.S. long-distance market, over $14 billion in revenues in 1998, and the dynamic changes that have been occurring in that segment, there has been surprisingly little research in this area. Moreover, much of the work that has been done has focused upon the economic determinants of international settlement rates[81]. We were able to locate only three studies dealing directly with the intensity of competition in this market segment.

The first such study is by Alleman, Madden and Savage (1999) (AMS). In this study, AMS utilize the methodology developed by Kahai, Kaserman and Mayo (1996) to test for the degree of market power held by AT&T in the provision of international long-distance calls originating in the United States. In particular, AMS posit that the market for such calls has been consistent with the dominant firm-competitive fringe model of competition. Drawing upon this framework, AMS utilize known market share values for AT&T, together with econometric estimates of the elasticity of supply of fringe firms and the market demand to generate estimates of the residual demand elasticity for international calling faced by AT&T.

Data for 1991–1995 were gathered for the estimation, and the supply equation and the market demand estimation are estimated simultaneously. The results

[81] Settlement rates are the per-minute prices that foreign carriers charge U.S. companies to terminate international calls. U.S. carriers, in turn, receive settlement payments from foreign carriers for international calls terminated in the U.S. Settlement rates historically have been quite high (some exceeding $1.00 per minute) and have varied widely across countries. See Maleug and Schwartz (2000), Blake and Lande (1999) and Wright (1999).

indicate fringe supply elasticities ranging from 5.4 to 12.4 and market demand elasticities ranging from −0.93 to −0.68. Together with market share data, AMS are able to compute Lerner Index values for AT&T ranging from 0.12 to 0.24. Given the theoretical upper bound of the Lerner Index (unity) and the benchmark observations of the Lerner Index in other industries[82], AMS conclude that AT&T does not hold substantial market power in the provision of international calling over the 1991–1995 period.

The second study of market structure, competition, and pricing in the international telephone services market is that of Madden and Savage (forthcoming). In this study, Madden and Savage examine the relationship between the pricing of international calls originating in the United States, the prices charged by 'home' countries for providing access services, known as 'settlement rates' when describing international calling, and market structure. Specifically, Madden and Savage develop a model of U.S. carrier pricing based on a single-shot two-stage game in which domestic carriers first collectively establish a U.S. settlement rate, then individually establish a prices for international calling that maximize individual firm profits.

The model is used to motivate the use of a supply-relation in the form:

$$P = MC + \tau q \tag{20}$$

where P is the price of originating international toll calls, MC is the marginal cost of the calls, q is output and τ is an indicator, (as in Bresnahan [1989]) of the competitiveness of the market. The parameter τ, in turn, is modeled as a function of: (1) the degree of market concentration for outgoing traffic; (2) a measure of competition at both the originating and terminating ends of the call; and (3) the extent of privatization of the carriers operating at both ends of the bilateral market. As τ tends toward zero, industry behavior is construed to be more consistent with competitive performance, while higher values of τ indicate the presence of greater market power. Because individual carrier data are not generally available, Madden and Savage rely upon aggregate, country to country, data for purposes of estimating the supply relation and, accordingly, interpret the estimate of t as applicable to the 'average' firm. Data for 39 bilateral markets from 1991–1994 form the basis of estimation.

In recognition of the endogeneity of price and quantity, Madden and Savage establish an econometric model with a system of four equations, an outgoing and incoming supply relation and both incoming and outgoing demand equations. With respect to the variables of most interest, Madden and Savage find that decreases in the share of the largest firm in the U.S. (outgoing) calling market leads to statistically significant decreases in the price of international calling.

[82] See Bresnahan (1989) and Kahai, Kaserman and Mayo (1998).

Madden and Savage also find that, as the number of foreign country facilities-based providers increases, the price of U.S. originating calls falls. The competitive mechanism by which increases in the number of foreign carriers in their 'home' country leads to decreases in U.S. carrier prices is somewhat elusive. In this regard, Maddens and Savage's use of foreign facilities-based carriers rather than 'home' facilities-based carriers misses an opportunity to directly test the (arguably more important) question of whether increases in the number of 'home' facilities-based carriers affects pricing. That is, does the number of U.S. facilities-based carriers affect the price of outgoing U.S. international calling?

Finally, the Madden and Savage results indicate that privatization by foreign countries is associated with lower prices for both calls originating from those foreign countries and for calls originating in the U.S. that terminate in those markets. This, of course, raises an interesting question about the source of the declines in U.S. outgoing call prices in response to foreign country privatization. Several possibilities, unexplored in the Madden and Savage paper, arise. For instance, is the U.S. originating call price decrease the direct result of declines in foreign country prices brought about by privatization? Alternatively, is there a more indirect influence? That is, are lower U.S. outgoing prices brought about by foreign privatization the consequence of lower settlement rates following privatization? Answers to these questions await additional research.

The third study to address the competitiveness of the international segment of the long-distance market is by Ford (2000). This study adopts the general, but not unequivocable, approach, discussed in Section 4 above, of evaluating the intensity of competition by estimating the percentage flow through of marginal cost changes to equilibrium prices in a market. Here, the issue is how observed differences in international settlement rates affect the prices charged by AT&T and MCI to their U.S. customers for calls placed to 147 countries in 1997.

In this paper, the relationship between country-specific settlement rates and U.S. carriers' marginal costs is modelled explicity. As Ford explains, most countries currently exchange international telecommunications traffic under the International Settlements Policy (ISP). This policy, in turn, specifies the practice known as 'proportionate returns,' under which each domestic carrier receives settlement payments from foreign carriers that are dependent upon: (1) that carrier's share of U.S. originating minutes flowing to the specific country; (2) the total number of terminating minutes in the U.S. received from that country; and (3) the settlement rate that applies to that country. Specifically, under this policy, the total settlement cost to the domestic carrier for traffic carried to country i is given by:

$$c_i = S_i q_i - S_i F_i(q_i/Q_i), \tag{21}$$

where c_i is total settlement cost associated with calls to country i, S_i is the settlement rate in country i, q_i is the number of minutes carried by this firm to

country i, F_i is the total number of terminating minutes received in the U.S. from country i, and Q_i is the total number of originating minutes in the U.S. that are terminated in country i.

Note that all variables in equation (21) are country specific. The first term on the right hand side represents the firm's settlement payments and the second term shows the firm's settlement receipts[83]. Due to the presence of the latter, the firm's costs are not changed proportionately when settlement rates change. As Ford (2000) points out, it is important to recognize this fact in conducting a flow-through analysis of settlement rate changes.

Given Equation (21), the domestic firm's marginal cost of carrying an additional minute of traffic to country i is:

$$MC = S[1 - F/Q(1 - w)] \tag{22}$$

where MC is marginal cost, and w is the firm's share of the total traffic carried from the U.S. to that country[84]. The relationship between marginal cost and a change in the settlement rate is then given by:

$$\partial MC/\partial S = 1 - F/Q(1 - w). \tag{23}$$

Note that, in most instances, $F < Q$ and $w < 1$[85]. As a result, $\partial MC/\partial S < 1$ will hold in all, or virtually all, cases.

Ford calculates the values of $\partial MC/\partial S$ from equation (23) for both AT&T and MCI for his sample of 147 countries. He then incorporates the resulting variables in separate regressions for each company. Nine regional dummy variables are added to control for non-settlement cost variations attributable to such factors as the mix of transmission facilities (cable versus satellite) and other potential cost determinants. Both OLS and instrumental variable results (to correct for potential endogeneity of the settlement cost variable) are reported.

The estimated equations appear to perform quite well, explaining 88 percent of the observed variation in prices and obtaining statistical significance at the 0.01 level for all coefficient estimates. The coefficients associated with the settlement cost variable range from 1.027 to 1.040, with t-values of 21.927 to 23.342, respectively. Thus, the international segment of the long-distance market appears to exhibit complete pass through of settlement cost changes.

Next, these empirical results are used to calculate implied conjectural variation and residual demand elasticities under the assumption that market demand

[83] Analysis of settlement rate flow through differs from the analysis of access charge flow through in that there are no offsetting receipts to consider in the latter.

[84] We have dropped the country-specific subscript for notational simplicity.

[85] U.S. firms typically originate more traffic to a given country than they terminate from that country. Also, $w < 1$ holds for all countries where the U.S. carrier is not a monopolist.

has a constant price elasticity of −1.0. From these calculations, Ford (2000) concludes:

> "Thus, the IMTS industry is far more competitive than the standard Cournot model would imply, and very close to perfect competition. These results suggest that domestic IMTS carriers face residual demand elasticities (on average) of about −2.0..." [Footnote omitted].

In summary, relatively few studies have emerged to date on the extent of competition in the international calling market. Those studies that have appeared suggest that competititive pressures have been significant in limiting the degree of market power exhibited by U.S. based carriers.

7. Conclusion

In light of the above evidence, what can we conclude about the intensity of competition in the U.S. long-distance market? Is this market competitive? Like so many questions in economics, the answers appear to be: it depends. Specifically, it depends upon what is meant by the term 'competitive.' That is, what standard or benchmark level of competition is being applied, or implied, in posing this question? Indeed, one is reminded of the well-known adage of George Stigler (1956).

> "To determine whether an industry is workably competitive, therefore, simply have a good graduate student write his dissertation on the industry and render a verdict. It is crucial to this test, of course, that no second graduate student be allowed to study the industry."

The problem here, of course, is that we do not have a single graduate student. Instead, we have a cadre of analysts, many of whom have represented telecommunications firms with opposing public policy interests. Consequently, the absence of a ubiquitous opinion hardly comes as a surprise.

If, by 'competition,' we mean that sufficient rivalry is present to yield positive net social benefits from replacing direct price regulation of any sort by the relatively unrestrained operation of market forces, a variant, perhaps, on the concept of workable or effective competition, then we believe that the cumulative evidence overwhelmingly supports an affirmative answer. In particular, the evidence we have surveyed strongly suggests that the long-distance market is at least as competitive as many unregulated markets that yield acceptable levels of economic performance. On the basis of that evidence, then, the long-distance market is competitive under this standard.

If, however, we are asking whether competition is so intense that consumers benefits would be unlikely to flow from any additional entry into the market, putting aside any potential anti-competitive consequences that might accompany such entry in either this or local exchange markets, then the answer is considerably less certain. The evidence that has been reported to show the presence of tacit collusion and incomplete pass through of access charge reductions is generally

unconvincing. Nevertheless, the simple fact is that few, if any, real world markets would not benefit from some additional entry, particularly in industries experiencing rapid technological change and a breakdown of prior market boundaries. In such industries, new entrants may bring new, lower cost methods of production, economies of scale and scope, and novel approaches to bundled service offerings that appeal to specific consumer groups' preferences. As a result, even markets that are highly competitive are likely to experience some improved performance from entry. The policy implications of this latter standard are unclear, however, at least regarding RBOC entry, because of the potential anti-competitive consequences that may accompany such entry, particularly in local exchange markets. Such consequences, however, are not within the purview of this chapter.

References

Alleman, J., G. Madden, and S. J. Savage, 1999, Dominant carrier market power in United States international telephone markets, Mimeo.

Andrews, E. L., 1995, Finding the best deals among long-distance calling plans, New York Times, January 21.

Areeda, P. and D. Turner, 1975, Predatory prices and related practices under section 2 of the Sherman Act, Harvard Law Review, 88, 697–733.

AT&T Bell Laboratories, 1995, An updated study of AT&T's competitors' capacity to absorb rapid demand growth, Unpublished Report.

Bain, J. S., 1956, Barriers to new competition, Cambridge, MA: Harvard University Press.

Beard, T. R., 1999, Pass through of cost changes: Theoretical issues and policy, Mimeo, Auburn University Policy Research Center.

Beard, T. R., D. L. Kaserman and J. W. Mayo, 1999, Monopoly leveraging, path dependency, and the case for a local competition threshold for RBOC entry into interLATA toll, in M. A. Crew, ed., Regulation under increasing competition, Boston: Kluwer Academic Publishers.

Blake, L. and J. Lande, 1999, Trends in the U.S. international telecommunications industry, Industry analysis division, Common Carrier Bureau, Federal Communications Commission, September.

Bresnahan, T., 1989, Empirical studies of industries with market power, in R. Schmalensee and R. D. Willig, eds. Handbook of Industrial Organization, Amsterdam: North Holland.

Brock, G. W., 1981, The telecommunications industry: The dynamics of market structure, Cambridge, MA: Harvard University Press.

Brock, G. W., 2002, Historical overview, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science B.V., 43–74.

Bulow, J. I. and P. Pfleiderer, 1983, A note on the effect of cost changes on prices, Journal of Political Economy, 91, 182–185.

Burton, M., D. L. Kaserman and J. W. Mayo, 1999, Modeling entry and barriers to entry: A test of alternative specifications, Antitrust Bulletin, Summer, 387–420.

Chamberlin, E. H., 1933, The theory of monopolistic competition, Cambridge, MA: Harvard University Press.

Crandall, R. W. and L. Waverman, 1995, Talk is cheap: The promise of regulatory reform in North American telecommunications, Washington DC: The Brookings Institution.

Department of Justice and Federal Trade Commission, 1992, Horizontal Merger Guidelines.

Edelman, S. A., 1997, The FCC and the decline in AT&T's long distance residential rates, 1980–1992: Did price caps do it? Review of Industrial Organization, 12, 537–553.

Faulhaber, G. R., 1975, Cross-subsidization: Pricing in public enterprises, American Economic Review, 65, 966–977.
Faulhaber, G., 1987, Telecommunications in turmoil: Technology and public policy, Cambridge, MA: Ballinger Publishing Company.
Federal Communications Commission, 1995, re Motion of AT&T corp. to be reclassified a non-dominant carrier order, CC Docket No. 79-252, FCC 95-437, October 23.
Ford, G. S., 2000, Flow-through and competition in the international message telephone service market, Mimeo.
Hall, R. E., 1988, The relation between price and marginal cost in U.S. industry, Journal of Political Economy, 96, 921–947.
Hause, J. C. and G. DuRietz, 1984, Entry, industry growth and the microdynamics of industry supply, Journal of Political Economy, 92, 733–757.
Hausman, J. A., 1995, Competition in long-distance and telecommunications equipment markets: Effects of the MFJ, Managerial and Decision Economics, 16, 365–383.
Hill, R. C. and T. R. Beard, 1999, A statistical analysis of the flow-through of reductions in switched access charges to residential long distance rates, Mimeo.
Hovenkamp, H., 1990, Antitrust analysis of market power, with some thoughts about regulated industries, in J. R. Allison and D. L. Thomas, eds., Telecommunications Deregulation: Market Power and Cost Allocation issues, Westport, CT: Quorum Books.
Kaestner, R. and B. Kahn, 1990, The effects of regulation and competition on the price of AT&T intrastate telephone service, Journal of Regulatory Economics, 2, 363–378.
Kahai, S., D. L. Kaserman and J. W. Mayo, 1995, Deregulation and predation in long-distance telecommunications: An empirical test, Antirust Bulletin, 40, 645–666.
Kahai, S., D. L. Kaserman and J. W. Mayo, 1996, Is the 'Dominant Firm' dominant? An empirical analysis of AT&T's market power, Journal of Law and Economics, 39, 499–517.
Kaserman, D. L. and J. W. Mayo, 1986, The ghosts of deregulated telecommunications: An essay by exorcists, Journal of Policy Analysis and Management, 6, 84–92.
Kaserman, D. L. and J. W. Mayo, 1990, Deregulation and market power criteria: An evaluation of state level telecommunications policy, in J. R. Allison and D. L. Thomas, eds., Telecommunications Deregulation: Market Power and Cost Allocation issues, Westport, CT: Quorum Books.
Kaserman, D. L. and J. W. Mayo, 1995, Government and business: The economics of antitrust and regulation, Fort Worth, TX: Dryden Press.
Kaserman, D. L. and J. W. Mayo, 1996, Competition and asymmetric regulation in long-distance telecommunication: An assessment of the evidence, CommLaw Conspectus, 4, 1–26.
Katz, M. L. and R. D. Willig, 1983, Toward competition in telephone service: The case for freeing AT&T, Regulation, July/August, 43–49.
Klemperer, P., 1987, Entry deterrence in markets with consumer switching costs, Economic Journal, 97, 99–117.
Knittel, C. R., 1997, Interstate long distance rates: Search costs, switching costs, and market power, Review of Industrial Organization, 12, 519–536.
Landes, W. M. and R. Posner, 1981, Market power in antitrust cases, Harvard Law Review, 94, 937–996.
MacAvoy, P. A., 1995, Tacit collusion under regulation in the pricing of interstate long-distance telephone services, Journal of Economics and Management Strategy, 4, 147–185.
MacAvoy, P. A., 1996, The failure of antitrust and regulation to establish competition in long-distance telephone services, Cambridge, MA and Washington, D.C: The MIT Press and the AEI Press,
Madden, G. and S. J. Savage, 2001, Market structure, competition and pricing in United States international telephone markets, Review of Economics and Statistics, forthcoming.
Malueg, D. and M. Schwartz, 2000, Asymmetric telecom liberalization and gaming of international settlements via carrier alliances, Working Paper, Georgetown University.

Mathios, A. D. and R. P. Rogers, 1989, The impact of alternative forms of state regulation of AT&T on direct-dial long-distance telephone rates, RAND Journal of Economics, 20, 437–453.
McCormick, R. E., W. F. Shughart II and R. D. Tollison, 1984, The disinterest in deregulation, American Economic Review, 74, 1075–1079.
Mitchell, B. M. and I. Vogelsang, 1991, Telecommunications pricing: Theory and practice, Cambridge: Cambridge University Press.
Peltzman, S., 1976, Toward a more general theory of regulation, Journal of Law and Economics, 19, 211–240.
Posner, R., 1976, Antitrust law: An economic perspective, Chicago: University of Chicago Press.
Saving, T. R., 1970, Concentration ratios and the degree of monopoly, International Economic Review, 11, 139–146.
Scherer, F. M. and D. Ross, 1990, Industrial market structure and economic performance, Chicago: Rand McNally, Second Edition.
Schmalensee, R., 1989, Inter-industry studies of structure and performance, in R. Schmalensee and R. D. Willig, eds., Handbook of Industrial Organization, Amsterdam: North Holland.
Schwartz, M., 2000, The economic logic for conditioning bell entry into long-distance on the prior opening of local markets, Journal of Regulatory Economics, 18, 3, 247–288.
Shepherd, W. G., 1995, Contestability vs. competition-once more, Land Economics, 71, 299–309.
Spence, A. M., 1977, Entry, capacity, investment and oligopolistic pricing, Bell Journal of Economics, 8, 534–544.
Staiger, R. W. and F. A. Wolak, 1992, Collusive pricing with capacity constraints in the presence of demand uncertainty, RAND Journal of Economics, 23, 203–220.
Stigler, G., 1956, Report on antitrust policy-discussion, American Economic Review, 46, 505.
Sumner, D. A., 1981, Measurement of monopoly behavior: An application to the cigarette industry, Journal of Political Economy, 89, 1010–1019.
Taylor, L. D., 1994, Telecommunications demand in theory and practice, Boston: Kluwer Academic Publishers.
Taylor, L. D., 2002, Customer demand analysis, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds. Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science B. V., 97–142.
Taylor, W. E. and L. D. Taylor, 1993, Post-divestiture long-distance competition in the United States, American Economics Review, 83, 185–190.
Taylor, W. E. and J. D. Zona, 1997, An analysis of the state of competition in long-distance telephone markets, Journal of Regulatory Economics, 11, 227–255.
Temin, P. with L. Galambos, 1987, The fall of the Bell system, Cambridge: Cambridge University Press.
Ward, M., 1999, Product substitutability and competition in long-distance telecommunications, Economic Inquiry, 37, 657–677.
Wright, J., 1999, International telecommunications, settlement rates, and the FCC, Journal of Regulatory Economics, 15, 267–291.

This Page Intentionally Left Blank

Chapter 13

MOBILE TELEPHONE

JERRY HAUSMAN*

Massachussetts Institute of Technology

Contents

*I completed this paper in August 2000. Changes continue to occur in the mobile industry.

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

Mobile (cellular) telephone is an example of a new product that has significantly affected how people live. Since its introduction in Europe and Japan in the early 1980s and in the U.S. in 1983, mobile telephone adoption has grown at about 25 percent-30 percent per year such that at year end 1999 about 86 million mobile telephones are in use in the U.S.[1] Approximately 32 percent of all Americans use cellular, and there are about 66 percent as many cellular telephones in the U.S. as regular (landline) telephones. Penetration rates in Europe are significantly higher, with the penetration in Finland exceeding 70 percent. The average U.S. cellular customer spends about $500 per year on cellular service. Thus, consumers and businesses have found mobile telephone to be a valuable addition to their life-styles. Mobile telephones have created a significant increase in consumer welfare. They are one of the most significant innovations in the past twenty years.

Cellular telephone has been in commercial operation in the U.S. for seventeen years. Cellular telephone began operation in Chicago in late 1983 and in Los Angeles during the 1984 Olympic Games. Operation then began within the next year in the top 30 MSAs (Metropolitan Statistical Areas) and subsequently spread to the rest of the approximately 300 MSAs and more recently the RSAs (Rural Statistical Areas). Mobile telephone, which includes cellular plus PCS, is now available almost everywhere within the United States.

Mobile telephone, along with 800 telephone service, has been the great success story of new telecommunications services offered in the past 40 years. At the time of the AT&T divestiture, when it was not clear whether AT&T or the divested Bell Operating Companies (BOCs) would inherit the cellular spectrum which the Federal Communications Commission (FCC) had granted to AT&T, an AT&T prediction for cellular subscription levels in the year 1999 was about 1 million[2]. At year-end 1999 planning horizon, cellular subscribership in the U.S. exceeded 86 million, so that the AT&T forecast was off by approximately 85 million. In Figure 1, the number of cellular subscribers is graphed, with the growth rate of 25–30 percent a year holding for the past decade. Beginning in 1996 the next generation cellular technology, PCS, was introduced in the U.S. so that in most locations consumers have a choice among 5 or more mobile providers. AT&T Wireless, Sprint PCS, Verizon, and Nextel all offer nationwide networks, with a new nationwide network being constructed by Voicestream[3]. BellSouth and SBC offer large regional networks while many companies offer local or regional services[4].

[1] Data are from Spring 2000 CTIA Semi-Annual Wireless Industry Survey. See http://www.wow-com.com/wirelesssurvey/.

[2] This prediction for AT&T was done by McKinsey, a well-known consulting company. The lost value to AT&T by giving the cellular licenses to the BOCs was approximately $50 billion.

[3] Verizon is the combination of the former Vodafone AirTouch and Bell Atlantic networks. Voicestream is the combination of the former Voicestream, Omnipoint, and Aerial PCS networks.

[4] BellSouth and SBC are considering combining their mobile networks.

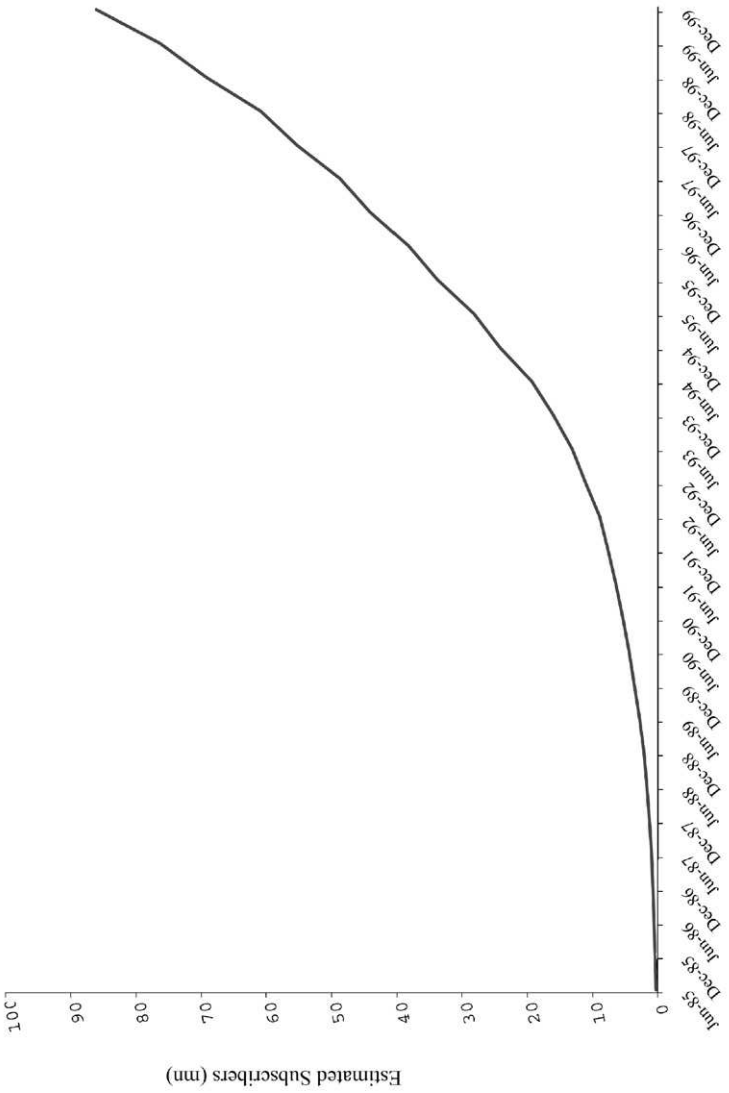


Fig. 1. Cellular subscribers in the U.S.

While mobile telephone has been a great success in the U.S., it has been even more successful in other countries. In Finland, Sweden, and Japan mobile penetration has now surpassed regular telephone (landline) penetration. In much of Europe, Japan and Australia mobile penetration is higher than in the U.S. The greater success in Europe and Japan of mobile is explained in part by the lack of subsidy and higher monthly price for local wireline telephone service that exists in the U.S[5]. Thus, mobile is relatively less expensive in these countries than in the U.S. Next, in Europe and Australia a single second generation digital mobile standard (GSM) was adopted, while in the U.S. three different standards are in use, which makes using a mobile telephone in non-local regions (called roaming) less convenient. Also, in Europe they have adopted a 'calling party pays' framework so incoming calls to mobile telephones are not charged to the recipient. In the U.S. the recipient pays, in a 'receiver pays' framework, so that a much lower proportion of mobile calls are incoming calls[6]. Lastly, in both Europe and Japan more innovative marketing has been used than in the U.S. Prepaid calling cards are have been highly successful in Europe, especially for young people, who need to control their budget[7]. In Japan, NTT DoCoMo has introduced Internet convenient mobile telephones that have a special appeal to young people[8]. Indeed, mobile companies worldwide are attempting to develop technology that will allow convenient combined use of mobile telephone and the Internet[9].

Given that mobile telephones and the Internet are the two outstanding telecommunications developments of the past 25 years, the question arises of the future of landline telephone. Modern digital mobile telephones are convenient to use, have good voice quality, are moderately secure, and the usage price is decreasing rapidly. For example, carriers offer monthly packages that cost approximately 10 cents to 15 cents per minute for combined local and long distance when large bundles of minutes are purchased. With the monthly subscription price at say $20 per month and long distance calls at approximately 7–9 cents per minute, landline telephone is still less expensive than mobile but lacks the convenience of 'anytime and anywhere,' especially the single number

[5] For instance in the U.S. residential local calls are free after paying a monthly fee while in Europe, Japan, and Australia local calls have a per minute or per call fee.

[6] In the U.S. many mobile subscribers do not give out their mobile telephone number because they have to pay for incoming calls. However, this policy seems to be changing as mobile subscriber use 'bucket' plans of buying a large number of minutes per month of mobile usage. I discuss these bucket plans below.

[7] In many European countries, e.g. Italy and the U.K., greater than 80 percent of new subscribers are using prepaid service. In the U.S. currently only about 6 percent of subscribers use prepaid service.

[8] In the first 14 months after its introduction NTT DoCoMo's I-mode telephone, which connects to the Internet, achieved approximately a 5 percent penetration among the Japanese population. Source: Wall Street Journal, April 17, 2000, p. A25.

[9] The leading technology at this time appears to be WAP (wireless access protocol) technology. A WAP browser will make Internet usage considerably easier for users.

capability of being reachable anytime and anywhere[10]. However, mobile prices do not have to decrease too much further before they could begin to have a significant substitution effect on landline telephone usage. The effect of mobile is already noticeable on payphone use, which has decreased by 17 percent over the past three years with significantly less usage and an absence of lines at airports.

However, in my view, it will be the Internet that will continue to provide an important role for landline service. Combined landline voice and DSL, which can provide high speed Internet access and combined cable Internet access and voice, will likely be used by customers to access the forthcoming broadband Internet. While next generation mobile (3G) may permit sufficient broadband access to the Internet, I believe that the shadow value of the spectrum, which is rationed by the FCC in the U.S. and other government agencies abroad, will limit its use. The value of the spectrum remains at a sufficiently high amount that mobile Internet access will not be widely used by residential customers except when they are 'on the go.' However, if government policy were to change from giving 'free usage' of the spectrum to other government bodies who do not use spectrum in an economical manner, the value of the spectrum could decrease sufficiently to make wireless the access method of choice ten years hence.

The average U.S. mobile telephone subscriber spends $41.24 per month on mobile service, or about $500 per year. Thus, about $41 billion per year is spent on mobile service, with additional amounts spent by consumers on purchasing mobile telephones. Mobile revenue now is about 40 percent as large as long distance revenue, so cellular telephone represents a significant expenditure category in telecommunications.

2. Description of the mobile industry

2.1. Technology

Cellular telephone technology introduced the concept of spectrum re-use within a given geographical area. Prior to the introduction of cellular telephone, car telephones made exclusive use of a given frequency band in a geographic area, which greatly limited the number of users. For instance in large U.S. cities such as New York and Los Angeles, IMTS service had waiting lists exceeding ten years[11]. Even for people that had service, repeated attempts to make a call often ended in failure because of system congestion.

[10] A typical large user makes about 500–600 minutes of local calls a month so the charge is still significantly less than mobile.

[11] IMTS stood for improved mobile telephone service.

Cellular technology, by allowing different users within the same geographic area, allowed for a tremendous expansion of capacity. Indeed, the early cellular networks were engineered to permit 95 percent of call attempts to be completed during peak usage periods. Cellular technology allowed for frequency reuse by using a hexagonal network of cells (See Figure 2A). Each base station covered one of the hexagons in Figure 2A. When a user went from one cell to an adjacent cell, the user would be handed off seamlessly to the new base station. Thus, non-adjacent cells would be using the same frequencies to serve cellular users[12]. In Figure 2C the cells have now been split to allow for increase frequency reuse at the cost of additional base stations and antennae. Lastly, in Figure 2B cell sectoring is displayed which uses three antennae, each transmitting over a 120-degree angle, replacing a single antenna which transmits over the full 360 degrees. The combination of cell splitting and sectorization increases capacity by approximately a factor of 8 times.

The original radio technology used in cellular was analogue technology, called AMPS in the U.S[13]. The AMPS system operated in the 800 MHz frequency range. European network operators originally began with AMPS-type systems usually in the 900 MHz range, but they agreed (with government help) to adopt subsequently a common digital standard, GSM. GSM was the first of the second-generation (2G) technologies, and it is the most widely used technology currently in the world for mobile telephones. GSM operates in the 900 MHz frequency range.

The U.S. did not adopt a common 2G technology. One group of companies, including AT&T Wireless, adopted TDMA, which is very closely related to GSM, although not interchangeable in use. GSM and TDMA have the advantage of increasing the number of users per frequency by approximately a factor of 3 over analog networks[14]. In the early 1990s the U.K. government decided to permit three new mobile companies to enter using frequency in the 1800 MHz range. Standard GSM technology was used, but the service was called PCN, for personal communications networks. The U.S. followed the U.K. lead in the mid-1990s by auctioning spectrum for up to 5 new entrants in the 1900 MHz frequency range. In the U.S. the technology is called PCS, for personal communications systems. GSM/TDMA is, by far, the most commonly used technology worldwide for cellular and PCS services.

The other primary 2G digital technology adopted in the U.S. is CDMA, for code division multiple access[15]. Rather than using a unique channel for a call, as

[12] Adjacent cells are not used because of potential interference problems.

[13] The switches to operate the system were often digital, which sometimes caused confusion in describing the systems. The analog systems were called TACS or NMT in Europe and operated in the 900 MHz or 450 MHz frequency bands.

[14] The increase in capacity is obtained by multiplexing voice messages so that 8 channels per frequency are permitted at a given time.

[15] CDMA is also called IS95.

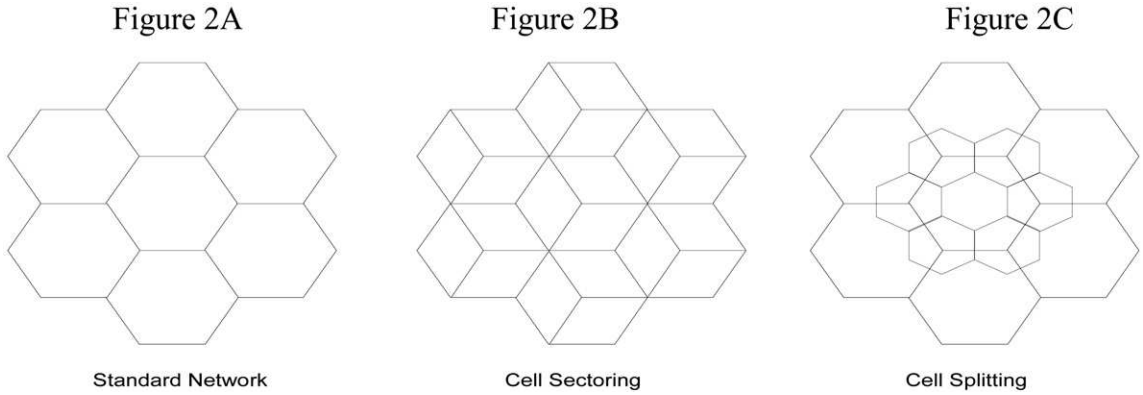


Fig. 2. Cell frequency design.

does GSM/TDMA, CDMA instead uses all the available frequencies for a given call. The technology is somewhat similar to packet switching in which a voice call (or data) is divided into individual packets, sent over available frequencies, and then reassembled for transmission to the other party. CDMA has greater capacity potential than GSM/TDMA with an increase of about 3 times over GSM[16]. CDMA also has superior handoff capability when a user transits from one cell to the adjacent cell as in Figure 2. When a GSM customer changes cells, typically the customer receives a new frequency channel. CDMA uses a 'soft' handoff when for a period of time the customer uses both the previous and future cell. This feature has increased importance in data transmissions where file transfers can become corrupted more easily with GSM. CDMA technology initially had substantial problems and significant doubts existed about its eventual success. However, these early problems have been overcome, and CDMA technology is now widely used[17]. In the U.S. the largest cellular operator, Verizon and also Sprint PCS have adopted CDMA technology.

Future third generation technology, often referred to as 3G, will have increased data usage at significantly higher speeds, rather than being primarily voice based networks[18]. Future networks will be constructed around an upgraded CDMA technology, given its technical superiority for data transmission[19]. One form of next generation technology, W-CDMA, is currently being used in Japan by NTT DoCoMo. To date its performance has been successful, with widespread usage and customer adoptions by DoCoMo, which offers an Internet based service. The other proposed approach, CDMA2000, proposes to offer significantly higher data speeds. This evolving technology will be combined with two other innovations, which will allow for more convenient Internet usage.

The first development is WAP, wireless application protocol. WAP will speed up the Internet Data using the WML language, in place of the standard HTML, which is used on the Internet. WML is approximately 2 times more byte-efficient that HTML. The other development is *Bluetooth*, which is technology that allows rapid data transfer from mobile devices to LANs and PCs over short distances, within about a ten-meter range currently. Thus, mobile users will be able to conveniently upload and download large amounts of data automatically, without having to use the mobile network to do so. While the replacement for cable technology and modems is obvious for *Bluetooth*, the lack of a 'killer application'

[16] The greater capacity might be pictured as using one 100 inch pipe rather than using ten 10 inch pipes to pass water through. Further, the problem of interference causing non-use of channels in adjacent cells is no longer a problem. Significant disputes exist about the actual increase in capacity that CDMA delivers over GSM.

[17] The resolution of the uncertainty is evident in the historical stock price and market capitalization of Qualcomm, the inventor of much of the CDMA technology.

[18] Most current mobile networks that are also used for voice do not allow for data transmission at speeds above 28.8 kbs, about 1/2 the speed of an analogue modem used with current PCs.

[19] Various intervening technologies, sometime called 2.5G, are currently being discussed and sometimes deployed. I will not discuss these interim technologies here.

to date may limit its acceptance. Personalized cash register information and vending machines with direct billing are much discussed future applications. Thus, the mobile telephone, or PDA (Palm), would permit an electronic payment system[20]. Also, *Bluetooth* might be useful to permit an intelligent personal phone number system that would give least cost routing depending on the customer location. Thus, calls would be routed over the landline system to a switch (PBX), which would transfer the call using *Bluetooth* to the mobile telephone without making use of the wireless network. This last application would greatly affect competition in telecommunications and in information networks. It would allow for 'edge networks' with most of the intelligence on the periphery of the network and with the mobile telephone used for communications and to integrate voice mail, email, internet information, and personalized transactions. This network design would be in sharp contrast to the centralized network model of the Internet, which has recently gained increased attention. Thus, the interaction of telecommunications and the Internet will greatly affect both network architecture and competition in the near future.

2.2. *Competing firms and countries*

Cellular technology was invented at Bell Labs (AT&T) in the 1960s and further developed by Motorola. These two firms, along with Ericsson, delivered the original analog cellular infrastructure, including base stations, radios, antennae, and switches (MSOs). However, neither AT&T nor Ericsson was successful in developing and selling handsets. Here Motorola took an early lead by incorporating modern electronics into its cellular telephones and introducing a very popular portable telephone. By the early 1990s the portable telephone surpassed 'car phones' because of its greater convenience and improvements in network technology that allow for high performance, even in automobiles. Motorola had well over 50 percent of worldwide sales of mobile telephones in the mid-1990s with Japanese and European manufacturers far behind.

In the late 1990s the competitive framework changed significantly. First, Nokia, a Finnish company that previously had gone from a forest products company to making televisions and PCs, became a major force in cellular handsets. Nokia was very successful based on its superior design approach to mobile handsets and their early embrace of digital technology, which Motorola was slower to concentrate on[21]. Nokia is now the leading producer of handsets, having surpassed Motorola[22]. In 1999 Nokia had approximate a 27 percent share of mobile telephone

[20] A PDA personal digital assistant allows for wireless transmission of data, but it does not have voice capability.

[21] Motorola faced the classic problem of being the world's most successful analog cellular company. Nokia did not have this initial position to defend.

[22] However, Nokia has been much more successful in GSM/TDMA technology, not in CDMA. A future danger exists that Nokia may have similar problems to Motorola.

handset sales followed by Motorola with 17 percent, Ericsson, with 11 percent, and Samsung with 6 percent[23]. Nokia's market capitalization exceeded $200 billion, about 4 times as large as Motorola.

The other significant technological change was the emergence of Qualcomm, a U.S. based company that had done significant development work on CDMA. Qualcomm claims to have a significant intellectual property position in CDMA, protected by patents. While significant doubt existed initially about the success of CDMA in real world settings, in the past 3 years these initial doubts have disappeared. Qualcomm originally manufactured handsets, but sold this division in 1999 year to Ericsson. As almost a pure technology company, Qualcomm's market value was near $60 billion (in 2000) and exceeded that of Motorola.

European governments have adopted a common standard, GSM, and Ericsson and Nokia have largely specialized in this technology. Thus, the Europeans have adopted a common technology standard, in contrast to the U.S., which has no common standard. The Korean government has encouraged its telecommunications hardware company to concentrate on CDMA, so they are likely to become a significant competitive factor in the near future. An interesting question is whether government intervention to create a common standard will lead to greater innovation and adoption due to network externalities, or whether the U.S. approach of technology competition will lead to greater innovation.

In terms of service providers Vodafone is now the world's largest with significant holding in the U.K., the U.S. (AirTouch/Pacific Telesis), Germany (Mannesmann), much of Europe, and Australia and New Zealand. Vodafone's market capitalization was approximately $215 billion, over two times the market capitalization of AT&T, historically the world leader in telecommunications. The market capitalizations of Sprint PCS and AT&T Wireless in the U.S. were approximately $20–40 billion. The gain for Sprint is especially noteworthy because its investment in spectrum and networks is probably less that 20 percent of its market capitalization[24]. Lastly, mobile networks have provided much of the growth for landline network operators such as the BOCs in the U.S. and BT in the U.K. The traditional landline business has very low growth compared to the rapid growth of mobile usage and revenue.

2.3. Government frequency allocation

The U.S. government initially permitted two competitive cellular companies in each geographic region to obtain cellular licenses. The entry of competition,

[23] Source of shares is Dataquest.

[24] MCI/Worldcom appears to have made a costly strategic mistake in 1995 when it dropped out of the PCS auctions, deciding that the spectrum prices were too high. Five years later it appears that mobile telephone may be an essential component to an integrated telecommunications company strategy. MCI/Worldcom attempted to purchase Sprint along with its PCS assets. However, the proposed merger was rejected by both U.S. and European antitrust regulators.

while not for the first time in U.S. telecommunications, was a departure from the usual practice in landline telephone where a single provider served a given geographic region[25]. In the U.S. one of the two frequency allocations was given to the local (landline) telephone provider. The other frequency allocation was originally designed to be given through comparative hearings – a 'beauty contest' in which Washington insiders, mainly lawyers, attempted to capture the maximum amount of future rents that were available. This procedure proved to be unworkable because so much rent was available that the procedure ground to a halt through legal maneuvering. Subsequently, (free) lotteries were held with the winner selling the lucky frequencies. A lucky dentist sold the frequencies for Cape Cod, a popular vacation spot in the U.S., for a reported $50 million.

Most European countries followed the U.S., although many initially allowed only the landline telephone provider to be awarded the cellular frequencies. Subsequently, most countries introduced competition with one or two 'additional providers. However, one important difference was that European and other nations typically awarded nationwide licenses, while the U.S. licenses were based on geographical areas the size of metropolitan areas, called MSAs. While the U.S. used a single analog AMPS technology, these regional networks led to higher transactions costs for 'roaming' customers' use of their mobile handsets outside their home territories[26]. However, with the current use of three incompatible technologies in the U.S. – TDMA, CDMA, and GSM – problems have arisen for roaming. Dual mode telephones are required which adds to the cost of service provision. Recently, nationwide networks have become increasingly important in the U.S., with Verizon, AT&T Wireless, Sprint PCS, Nextel, and Voicestream. Each network uses a common technology and allows for low cost roaming, since inter-network transactions costs are eliminated.

A major development in government frequency allocation occurred in the U.S. in 1993 when Congress, in need of a new revenue source, required the FCC to auction off spectrum, rather than using comparative hearing or other procedures[27]. Economists had long favored auctions, rather than a political process that awarded spectrum to the politically well connected. Auctions lead to the greatest value use of the limited resource since those users will bid the most. Thus, auctions lead to economic efficiency[28]. Furthermore, the government captures the rents associated with limited spectrum, rather than rents going to lawyers or lucky lottery winners.

The FCC adopted auctions for the initial PCS auctions of two 30 MHz bands in the U.S. in late 1994–95. The auctions were quite successful with approximately

[25] Earlier, competition in the U.S. existed in both paging and in car telephone service, IMTS.
[26] In the U.S. roaming revenues have been approximately 10–12 percent of overall mobile revenues.
[27] The U.S. government had a significant deficit at this time.
[28] I will not discuss auctions in detail. See the chapter by Cramton in this *Handbook*.

$6 billion raised in revenue[29]. A subsequent auction for an additional 30 MHz of spectrum was limited to small companies as a political favor, but it led to a large number of bankruptcies and continuing litigation[30]. Subsequently, unrestricted auctions have not encountered difficulties. A number of other countries have also adopted auctions, e.g. the U.K. and Australia have recently held successful auctions. The recently (April 2000) completed auction in the U.K. for 3G technology spectrum raised approximately $35 billion for the government[31]. Germany also recently (August 2000) completed its auction for 3G spectrum that raised approximately $46 billion[32]. Evidently, the future promise of combined mobile telephones with the Internet has led firms to expect an explosion of use when 3G technology becomes available.

Auctions have demonstrated their superiority as an allocation mechanism for scarce spectrum allocation[33]. Thus, a number of other countries such as Italy, the Netherlands, Austria, and Belgium plan to auction off their 3G spectrum. However, France has decided not to renounce its Colbertian tradition and adopt a market allocation device. Instead, the French government will decide who the 'worthy' recipients of the 3G licenses will be. Perhaps, more surprisingly, the Scandinavian countries along with Spain and Portugal have decided to continue to use beauty contests[34]. Japan, Korea, and Hong Kong also plan to use beauty contests. The most-cited reason (at least in public) for the continued use of beauty contest spectrum allocation is because the large amounts paid by successful bidder will lead to high prices to consumers. This claim makes no economic sense unless future mobile prices to consumers will be regulated. Future demand, cost, and competitive factors will determine the service prices, with the license fees a sunk cost to the original winning bidders. Indeed, using the argument cited above, the European governments would soon be expected to begin beauty contests for the allocation of European football players so that the ticket prices would decrease to consumers.

[29] Actually, the FCC could not quite let go of its previous behavior. It awarded 3 "pioneer preference" licenses in New York, Los Angeles, and Washington at discounted prices. These awards were political giveways, e.g. the Washington Post, the dominant newspaper in the nation's capital, received one of the three pioneer preference awards. The award followed a dinner between President Clinton and the publisher of the Washington Post on Martha's Vineyard in August 1994.

[30] The FCC seems unable to stop its decades long practice of attempting to gain political favor by interfering in the efficient operation of the U.S. telecommunications networks.

[31] The U.K. government reserved one of five frequency bands for a new entrant.

[32] An interesting outcome is that the per capita auction amounts for the U.K. and Germany are extremely close, with the U.K. slightly higher.

[33] Potential problems of coordinated interaction among bidders have not occurred.

[34] A problem with beauty contests is choice of the 'best' technology. In December 2000, Telia, the largest mobile operator in Sweden, did not receive a 3G license. Telia announced it would appeal the decision by the government regulator. The regulator rejected Telia's bid on the ground of technical feasibility because it would not provide sufficient geographical coverage. Telia claims that the regulator made a technical mistake in evaluating its technology and rejected Telia's 3G bid "on false assumptions."

Another question that arises is inefficient spectrum use promoted by governments. The high value to spectrum arises because of spectrum scarcity. Much of the scarce spectrum is used by government agencies (both national and local) and the military, none of which have any economic incentive to use their spectrum efficiently. Given the modern technology that has been adopted by the mobile industry, these government agencies could also more economically utilize their spectrum. Governments have not yet fully realized the tens of billions (or hundreds of billions) of dollars of economic loss arising from their outdated spectrum policies. Government agencies should be required to bid for spectrum usage so that they will have the correct economic incentives for efficient usage. Given that the success of the future combination of mobile and Internet usage, for the broadband Internet, may well depend on sufficient spectrum being available, governments need to revisit their overall policies of spectrum usage. Spectrum policy is too important to be left in the hands of the FCC and similar government agencies, where politics typically is more important than economic efficiency or consumer welfare in making allocation and taxation decisions[35].

3. International comparisons

3.1. Performance across countries

Prices of mobile service are quite difficult to compare across countries because of currency fluctuations, differences in taxation, and differences in interconnection fees that cellular companies must pay landline telephone companies[36]. Thus, most inter-country comparisons focus on penetration[37]. However, as I noted above, penetration comparisons depend, in part, on the relative price of landline telephone service that varies greatly across countries.

In Figure 3, the graph shows recent (2000) mobile penetration levels for a sample of European countries. As previously noted, mobile penetration is particularly high in the Scandinavian countries—about 65 percent in each country[38]. In urban areas it is even higher, with estimates for Helsinki above 80 percent. Note that the U.S. penetration at about 35 percent is in the middle of the European penetration rates, around the level of the U.K. Thus, while the U.S. penetration levels are lower than many European countries, penetration is not closely related to income. Nor does the absence of 'calling party pays' in the U.S. nor the lack of a uniform national standard explain the differences between

[35] See e.g. Hausman and Shelanski (1999), where we estimate the loss of economic efficiency and consumers welfare to be in the billions of dollars per year because of recent FCC taxation policy.

[36] In the U.S. over 80 percent of cellular calls terminate on the landline network. In European countries the percentage is lower due to higher mobile penetration.

[37] Ahn and Lee (1999) conduct an econometric study of penetration across 64 countries.

[38] For a study of the determinants of mobile penetration in Europe see Gruber and Verboven (1999).

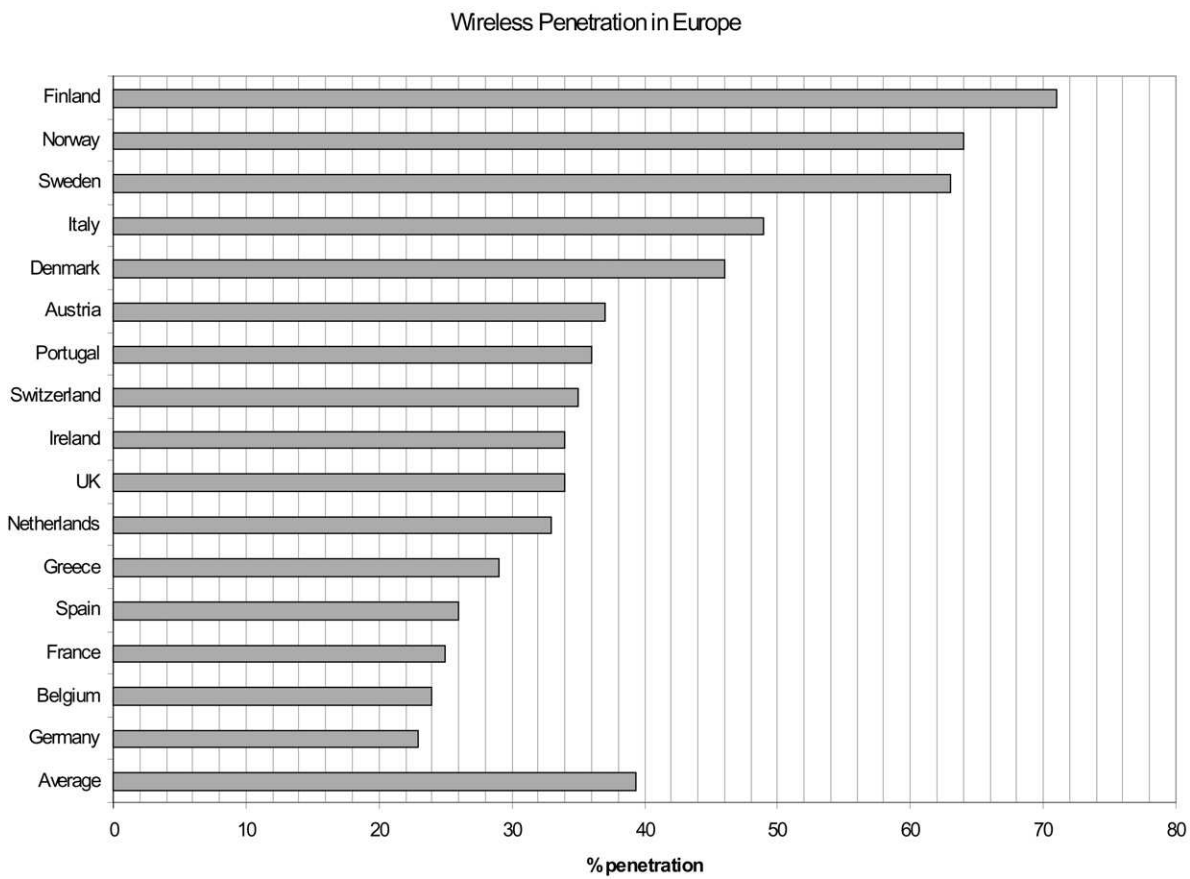


Fig. 3. Mobile penetration in Europe (2000).

Europe and the U.S[39]. Australian and Japanese penetration levels are above U.S. penetration levels and at around the European average penetration level of about 40 percent.

However, even high penetration ratios do not demonstrate that mobile has replaced landline to a large degree. Finland's penetration is above 70 percent, which far exceeds landline telephone penetration. Nevertheless in Finland mobile traffic only accounts for around 15 percent of total voice traffic in the Finnish network. I expect this number to grow over time as the 'younger generation' grows up using mobile as their primary choice. However, for now landline telephones still carry the large bulk of voice traffic in all countries[40].

Figure 4 presents price data collected by the OECD, which takes account of both monthly service price and handset prices. Handset subsidies are an important competitive factor since consumers place a great weight on the initial cost of cellular. Handset subsidies first became widely used in the late 1980 s in the U.S. and helped cause the rapid growth of cellular, after a rather slow start to consumer adoption[41]. Empirical research has demonstrated repeatedly that consumers pay 'too much' attention to the initial cost compared to the operating cost of a durable good. In the mobile context, competition has led to large discounts and subsidies for mobile handsets, as demanded by consumers. This outcome has been observed in Australia, the U.S., Canada, and the U.K. Mobile consumers are more likely to buy the service if the up-front handset cost is below the full (standalone) competitive price. Mobile companies' consumer research in the U.S. and Australia demonstrates that customers are most price sensitive to the up front costs of the price of handsets and monthly rental, which is consistent with market outcomes and my previous academic research[42].

As Figure 4 demonstrates, the U.S. has among the lowest total cost of mobile service. Among Scandinavian countries, Norway and Denmark are quite low, while Sweden and Finland are higher than the U.S. or Canada. Germany, which has relatively low penetration by European standards, does not have a particularly high price, while France has by far the highest total cost of mobile service. Thus, penetration is not particularly closely linked to prices charged to consumers[43].

[39] Years since launch of the system has some effect on penetration rates, but its influence is decreasing over time, as expected.

[40] For a recent study on the substitution of mobile for fixed calls in Portugal, see Barros and Cadima (2000).

[41] The initial cost of a service has been noted in many consumer decisions. I first discovered its importance in consumer purchasing decisions regarding energy efficient appliances. See Hausman (1979). Many subsequent papers have found the same result over a wide range of consumer durables. Consumers behave as if they had high discount rates, placing a premium on upfront costs.

[42] An alternative explanation of handset subsidies is to build market share in the presence of switching costs. See Valletti (1999b).

[43] Jha and Majumdar (1999) find an important role for cellular technology diffusion on the overall competitiveness of the telecommunications sector for 23 OECD countries over the period 1980–1995.

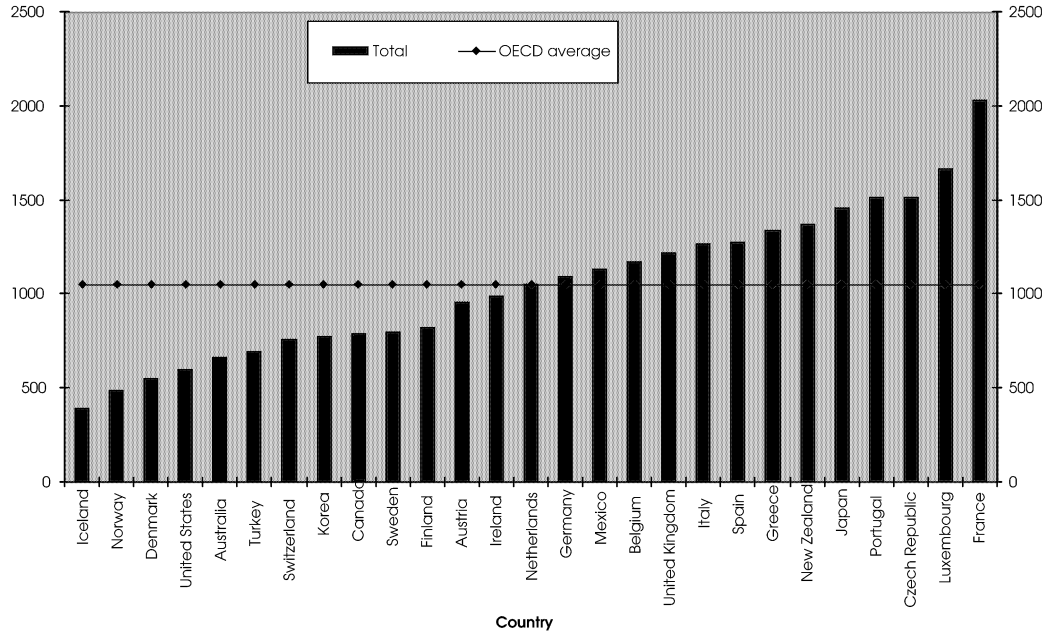


Fig. 4. Total cost of mobile service.
Source: OECD Bench marking "Communications Outlook" 1999.

3.2. Pricing within the U.S.

Up through 1996 mobile telephone in the U.S. operated as a cellular duopoly with two carriers in each region[44]. Beginning in 1996 two new mobile PCS carriers entered in most large MSAs, with additional entry since that time[45]. In Hausman (1999a), to construct a price index for cellular service, I used survey data from cellular companies in the top 30 MSAs since 1985. The top 30 MSAs comprise about 41 percent of the total U.S. population. To account for changes in cellular usage, the lowest price subscriber plan was calculated for 160 minutes per month of cellular usage with 80 percent peak and 20 percent off-peak usage. this plan was then used to compute a price index. In Figure 5, the price index for cellular service is graphed, with a base of 100 in 1985.

By the beginning of the year 1998 the index had decreased to 0.73 so that cellular prices had decreased by about 27 percent over the period, a decrease of about 2.6 percent per year. Note that price fell significantly in 1995–96 when the new entry of PCS occurred. Thus, as expected, new entry along with deregulation of prices by the FCC led to a faster decrease in prices than had previously occurred[46]. Combined with the significant decrease in retail cellular telephone prices, which is graphed in Figure 6, an overall index of the cost of cellular service and the cost of cellular telephones is graphed in Figure 7 where the overall decrease from the base of 100 in 1985 is 51 percent[47]. Thus, including both equipment and service prices the total cost of cellular decreased by about 5.8 percent per year over the period 1985–1997.

In April 1999 cellular pricing changed greatly in the U.S. Dan Hesse, president of AT&T Wireless, introduced an 'anywhere and anytime' (digital) one rate pricing plan. Thus, neither roaming nor long distance charges were assessed, and the customer paid a single rate wherever the call was placed and regardless of where (in the U.S.) the call terminated. Furthermore, AT&T and other carriers, such as Sprint, adopted 'bucket plans' where the customer bought a monthly bucket of minutes. The per minute prices are usually between 10 cents to 15 cents per minute, with lower prices arising with larger of buckets of minutes purchased. Most of the large mobile companies have since adopted these plans, which are by far the most widely advertised and sold plans.

[44] Valetti (1999a) has a model in competition among coverage areas for product differentiation. Here I concentrate on price competition, which seems to be the major focus of competition.

[45] This time period is approximately when the Nextel network, using "iden" technology, became fully operational.

[46] State regulation of cellular prices in approximately half of the states had led to cellular prices about 15 percent higher in the regulated states. See Hausman (1995b). FCC deregulation of cellular prices led to prices dropping in previously regulated states by significantly more than they decreased in previously non-regulated states.

[47] To create the combined index I assume a 'churn' (customer drop) rate of 0.33, which has been the approximate rate for the past decade.

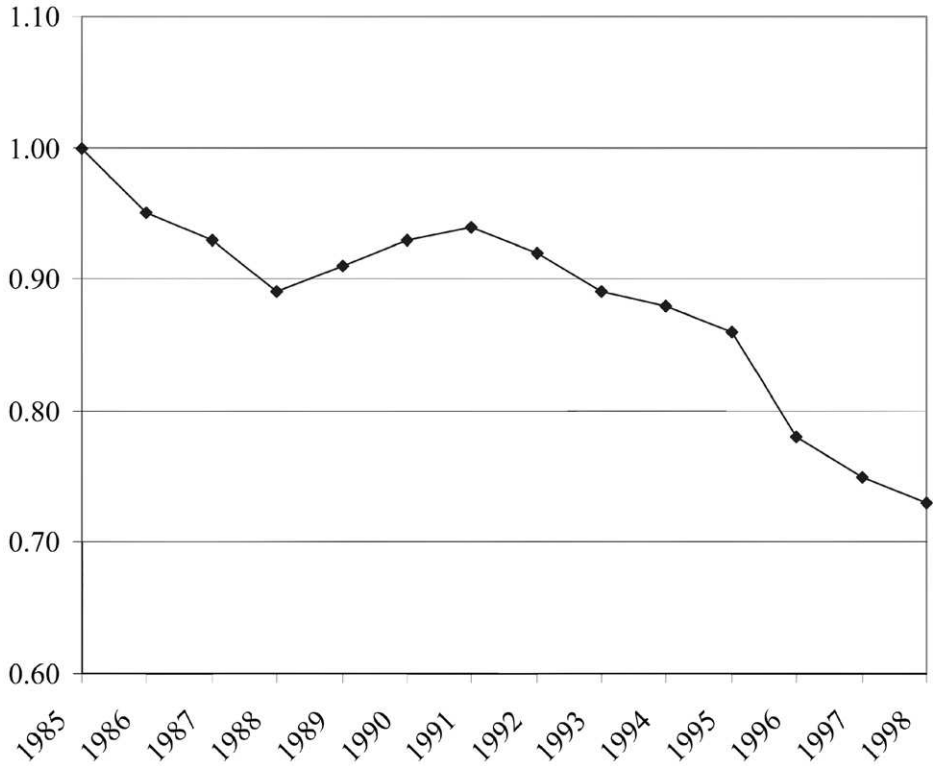


Fig. 5. Cellular average price index for top 30 MSAs: 1985–1998 (Base: 1985 = 100).

The first of the new PCS services began in late 1995, based upon the first two 30 MHz blocks of spectrum that were auctioned between December 1994 and March 1995. Since that time, a large number of companies have begun building facilities. All but 28 of the Top 100 metropolitan markets in the U.S. now have at least five wireless competitors: two cellular providers, two to four PCS services, and Nextel. Thus, about 70 percent of the U.S. population has a choice among five or more competing wireless carriers. The effect of the resulting competition on wireless rates in the U.S. has been significant. Throughout the 1984–1995 period real, inflation-adjusted cellular rates had fallen at a rate of 4 percent-5 percent percent per year[48]. Between 1995 and 1999, however, real cellular rates fell at a rate of 17 percent per year as PCS service providers offered service at prices per

[48] See Hausman (1999a).

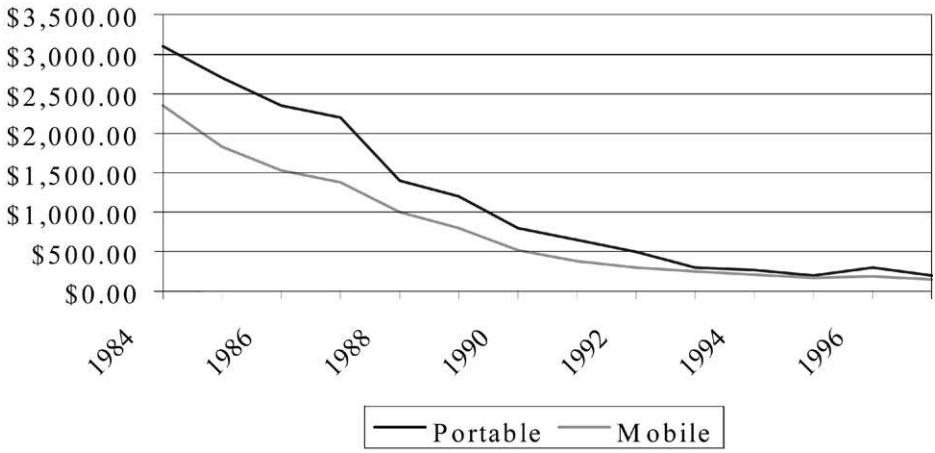


Fig. 6. Cellular telephone average prices: 1984–1997.

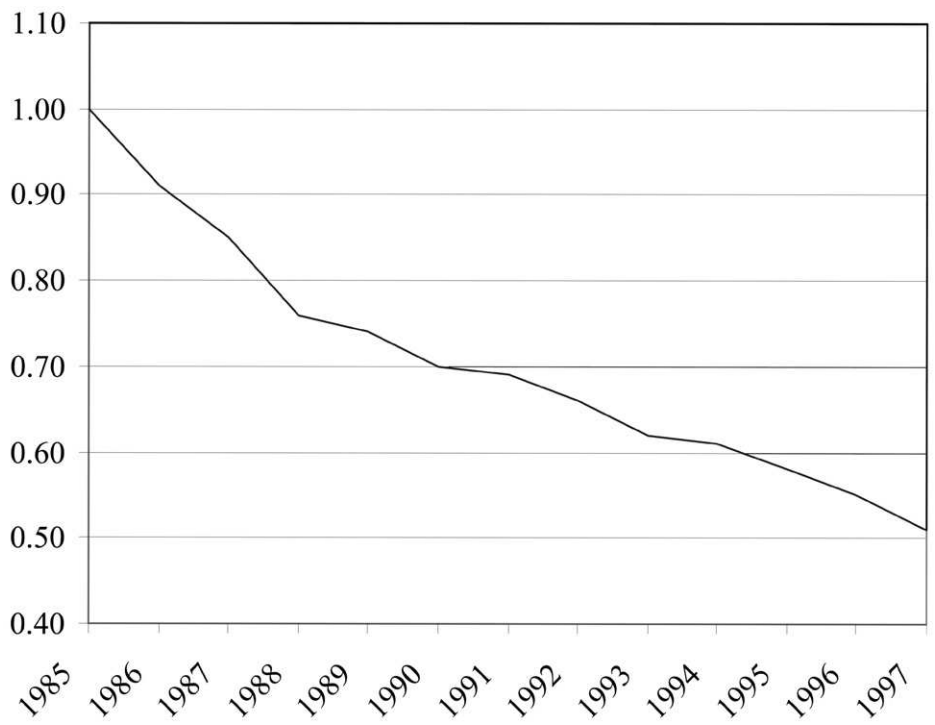


Fig. 7. Combined cellular service and equipment prices: 1985–1997 (Base: 1985 = 1.00).

minute in bucket plans that were more than 50 percent lower than existing cellular rates[49]. The U.S. Bureau of Labor Statistics reports that mobile service price decreased by 11.3 percent in the 12 months ending in January 2000. Crandall and Hausman (2000) conclude that the evidence shows that with open entry, only one new operator is required to drive rates sharply lower. Subsequent entrants have no further statistically significant effect on rates. Of course, the presence of nationwide carriers may be enough to cause mobile rates to react on a nationwide basis to the presence of an additional nationwide carrier, e.g. AT&T.

3.3. Pricing in Europe

The path of prices in Europe is more difficult to determine because of the lack of government price indices or academic studies. Nattermann (1999) calculates price changes in Germany over the period 1992–1998. He chooses the beginning of the period when competition began in the German cellular market with the entry of Mannesmann (in a joint venture with AirTouch) and the government eliminated price regulation. He estimates that a market share weighted price index decreased by 45.7 percent for a change of -7.4 percent per year. This price decrease is greater than the duopoly period in the U.S. but less than the price decreases since the entry of PCS operators in the U.S.

Valletti and Cave (1998) analyze competition in the U.K.. They divide the period from 1985 into the initial duopoly period with BT Cellnet and Vodafone and the subsequent period since the entry of the two additional PCN (PCS) operators One2One and Orange. In the initial duopoly phase of competition Valletti and Cave report that retail prices were extremely stable with few packages designed for consumers[50]. Valletti and Cave posit monopoly-like behavior of the duopolists, but they do not discuss the effect of regulation in causing this behavior.

However, with the entry of the PCN operators, competition increased significantly. The number of different pricing options increased greatly. Demand also increased significantly with lower prices and increasing choice of pricing plans. Unfortunately, Valleti and Cave do not do a systematic study of price changes or construct a price index, so the extent of price changes cannot be determined. However, they do give a list of price changes over time, which demonstrates that prices did decrease significantly after the entry of the two additional PCN operators.

[49] This estimate is consistent with an alternative estimate of a price decrease of 20 percent between 1998 and 1999. See FCC, "Fifth Report," August 18, 2000, p. 5. For comparison, in Australia Vodafone prices for an average mobile user decreased 60 per cent between 1993 and 1999, or about 8.2 percent per year, including access and handsets when 3 firms were competing. In the year after new network entry began to occur in Australia, mobile prices, including both access and handset prices, decreased by about 15 percent, which is consistent with the U.S. experience.

[50] Oftel, the telecommunications regulator, did not permit highly discounted handset prices as occurred in the US during this period.

4. Increased consumer welfare from mobile telephone

The introduction claimed that mobile telephone had created a significant increase in consumer welfare. I now measure this increased consumer welfare effect in the U.S. Two equivalent methods exist in economics to measure the increased consumer welfare: use of an expenditure function (consumers surplus) or estimation of a cost of living index (COLI). Both the expenditure function approach and the cost of living index approach answer the following question. How much more (or less) income does a representative consumer require to be as well off in period 1 as in period 0 given changes in prices, changes in the quality of goods, and the introduction of new goods (or the disappearance of existing goods)? I first explain the theory of cost-of-living indices and demonstrate how new goods should be included using the classical theory of Hicks (1940) and Rothbarth (1941) and then estimate the effect on consumer welfare[51].

4.1. Economic theory to measure increased consumer welfare

In either the expenditure function or the cost of living approach, we need to know what the price would have been in the pre-introduction period. The correct price to use for the new good in the pre-introduction period is the 'virtual' price, which sets demand to zero. Estimation of this virtual price requires estimation of a demand function that in turn provides the expenditure function, which allows exact calculation of the change in consumer welfare or the COLI. Given the demand function one can solve for the virtual price and for the expenditure function (or indirect utility function) and make correct evaluations of consumer welfare and the change in the COLI from the introduction of a new product or service. In period 1 consider the demand for the new good, x_n, as a function of all prices and income, y:

$$x_n = g(p_1, \ldots, p_{n-1}, p_n, y). \tag{1}$$

Now, if the good were not available in period 0, I solve for the virtual price, p_n^*, which causes the demand for the new good to be equal to zero:

$$0 = x_n = g(p_1, \ldots, p_{n-1}, p_n^*, y). \tag{2}$$

However, instead of using the Marshallian demand curve approach of Hicks (1940) and Rothbarth (1941) in Equations (1) and (2), I instead would use the income compensated and utility constant Hicksian demand curve to do an exact welfare evaluation. In Equation (2) income, y, is solved in terms of the utility level u^1 to find the Hicksian demand curve given the Marshallian demand curve specification.

[51] This section is based on Hausman (1999a).

In terms of the expenditure function one can solve the differential equation from Roy's identity which corresponds to the demand function in Equation (1) to find the (partial) expenditure function, using the techniques that was developed in Hausman (1981). The approach solves the differential equation that arises from Roy's identity in the case of common parametric specifications of demand:

$$y = e(p_1, \ldots, p_{n-1}, p_n, u^1). \tag{3}$$

The expenditure function gives the minimum amount of income, y, to achieve the level of utility u^1, which arises from the indirect utility function, which corresponds to the demand function of Equation (1) and the expenditure function of Equation (3). To solve for the amount of income needed to achieve utility level u^1 in the absence of the new good, I use the expenditure function from Equation (3) to calculate:

$$y^* = e(p_1, \ldots, p_{n-1}, p_n^*, u^1). \tag{4}$$

The exact cost of living index becomes $P(p, p^*, u^1) = y^*/y$. As with any index number calculation, one could use the pre-introduction utility level u^0 to calculate the cost of living index. However, almost no change would occur in y^*/y because of the relatively small percentage of expenditure on cellular telephone compared to income y.

Note that to use this approach one must estimate a demand curve as in Equation (1), which in turn implies the expenditure function and the ability to do the exact welfare calculation of Equations (3) and (4). Thus, the only required assumption is to specify a parametric form of the demand function[52].

Estimation of the expenditure function in Equation (4) as well as y^* typically requires significant amounts of data and estimation of a demand curve. Furthermore, estimation of the virtual price requires an estimate of the price at which demand equals zero. These data are typically unobservable and require extrapolation from the estimated demand curve. I also use an alternative conservative approach, which decreases the information requirements and provides a 'lower bound' estimate. Once the demand curve is estimated, an approximation can be used by taking the supporting hyperplane at the observed price and quantities, (p_1, q_1), which then leads to an estimate of the virtual price of the lower bound linear demand curve to the actual demand curve. Now I claim that this estimate is conservative because the estimated virtual price from the linear demand curve will be less than the virtual price from the actual demand curve, unless the 'true' demand curve is concave to the origin, which while theoretically possible would not be expected to occur for most new products and services. The change in

[52] A non-parametric approach to the problem could be used with techniques developed in Hausman and Newey (1995).

expenditure to hold utility constant with the introduction of the new product $y-y^*$ is the compensating variation which again can be approximated by the area under the approximate demand curve above the observed price. This amount is easily computed as:

$$y - y^* \approx CV = (0.5\ p_1 q_1)/\alpha \tag{5}$$

where CV is the compensating variation (consumers surplus) from the introduction of the new product and α is the own price elasticity of demand. To estimate Equation (5) current revenue $R = p_1 q_1$ is required. Here data from an industry source is used, the CTIA, which collects the data on a semi-annual basis. The only econometric estimate needed is for the price elasticity α, and this parameter appears to be the irreducible feature of the demand curve that is needed to estimate the change in the COLI from the introduction of a new product or service.

4.2. Estimation of the amount of increased consumer welfare

Now turn to the econometric estimation to implement the expenditure function approach of Equations (3) and (4) and the approximation approach of Equation (5). To do so, I collected price and subscribership data for the period 1989–93 from a (confidential) survey of cellular operators. These five years of data were used to run a log-linear regression of cellular prices in the top 30 MSAs. These top 30 MSAs contain about 107 million pops (population), or about 41 percent of the entire U.S. population. The estimated demand curve is reported in Hausman (1997). Using instrumental variables methodology to take account of possible joint endogeneity of price and demand, I estimate the demand elasticity to be -0.51 (standard error $= 0.17$). The results seem reasonable because the lower bound estimate (from the approximate linear demand curve) of the virtual price is \$97 per month, while the data set demonstrates that some actual monthly fees were as high as \$125 per month with significant demand at this price level. Thus, the conservative estimate of the lower bound virtual price does turn out to be quite conservative.

To calculate the expenditure function of Equation (4) the results of Hausman (1981) are used to calculate

$$e(p, \bar{u}) = [(1-\delta)(\bar{u} + Ap^{1+\alpha}/(1+\alpha))]^{1/(1-\delta)} \tag{6}$$

where A is the intercept of the demand curve, α is the price elasticity, and δ is the income elasticity estimate. The compensating variation is calculated from Equation (6) where y is income:

$$CV = \left\{ \frac{(1-\delta)}{(1+\alpha)} y^{-\delta}[p_1 x_1 - p_0 x_0] + y^{(1-\delta)} \right\}^{1/(1-\delta)} - y. \tag{7}$$

I then use Equation (7) to calculate the CV for the introduction of cellular telephone using the average revenue and subscribership data from the CTIA, which was discussed earlier, as well as the econometric estimates of the parameters of the demand function and associated expenditure function. I estimate that increased consumer welfare from cellular telephone in 1994 using Equation (7) was $49.8 billion, while using the more conservative approach of Equation (5) the amount was $24.2 billion. Using year-end 1999 data from the CTIA corresponding amounts of $111 billion and $52.8 billion were calculated. Thus, the average user of mobile telephone gains an increase in consumer surplus approximately equal to the yearly expenditure on the service.

In Figure 8, resulting price index for a COLI is graphed based on these results. Note that using a basis of 1988, the COLI index for cellular telephone decreases to about 0.116 by 1998. The calculated price index is extremely similar, whether the exact or approximate demand curve estimates are used. Note that the index of 0.116 is well below the price index, which is calculated in Section 2 to be 0.73. The index correctly captures the overall gain in consumer welfare from the introduction of cellular telephones.

4.3. Estimation of a corrected telecommunications services CPI

I now make a further approach to estimating the gain in consumer welfare from mobile telephone by re-estimating the telecommunications CPI in the U.S. to take account of mobile telephone. The BLS did not include mobile telephone in the telecommunications CPI until 1998 so that I recalculate a corrected telecommunications CPI that accounts for the effect of mobile telephones[53].

The BLS calculates a telephone services CPI each month. The three components of the telephone services CPI are local access charges, intrastate long distance (toll) charges, and interstate long distance (toll) charges. Cellular telephone had not been included in the telephone services CPI until 1998. The telephone services CPI is approximately 1.7 percent of the overall CPI.

I now estimate a corrected telecommunications services CPI where cellular service is included. I take into account the decline in cellular service prices, e.g. Figure 5, and also the gain in consumer welfare from the introduction of cellular service, e.g. Figure 8. Thus, a COLI based telecommunications services CPI is approximated.

To construct the corrected telecommunications services CPI yearly expenditures weights are used, based on total local and long distance expenditure for residential customers. Thus, I am estimating the welfare effect on consumers from cellular telephone, and eliminating the effect of business users, by assuming that consumers use cellular in the same proportion that consumers use long distance

[53] When the BLS introduced mobile into the CPI in 1998, it made no correction for previous decreases that would have affected earlier values of the CPI.

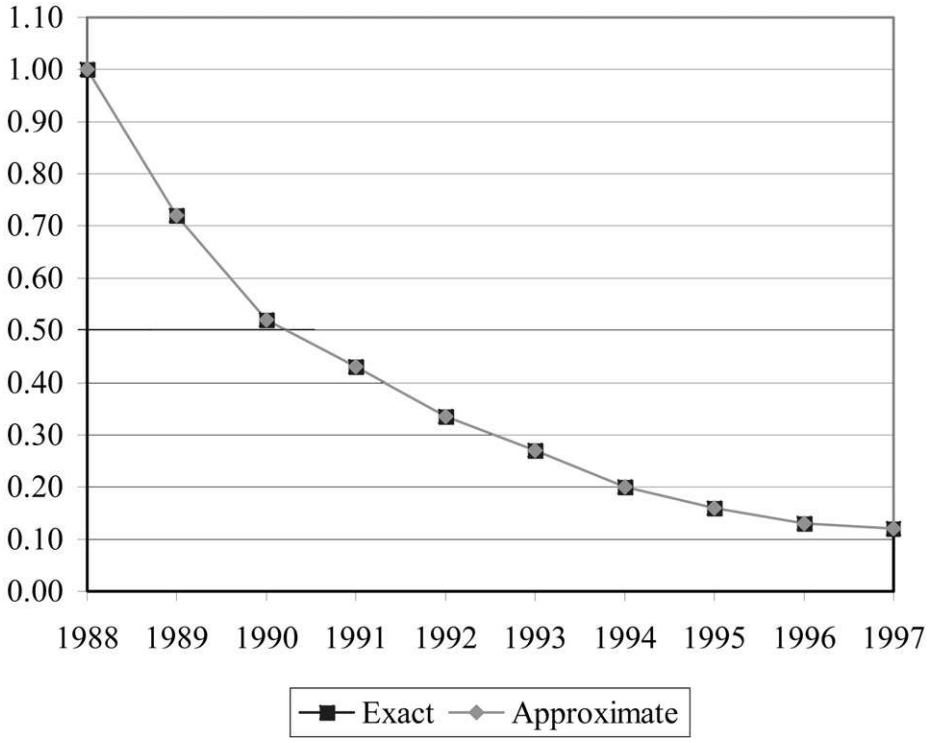


Fig. 8. Exact and approximate cost of living index.

service, compared to businesses. Since there is no expenditure share data across the various years, this method leads to an approximation. However, the approximation should be relatively accurate given the average expenditure of households on long distance with similar demographic characteristics to consumer cellular users. To the extent that the proportion of consumer usage of cellular is approximately equal to consumer usage of local and long distance services, these weights create a superlative price index[54]. Otherwise, the calculation leads to an approximation to a telecommunications CPI, which would need data on consumer expenditure shares to become a superlative index.

In Figure 9 the BLS telecommunications services CPI along with the corrected COLIs for telecommunications services is graphed. The price index from 'early inclusion' of cellular into the CPI in 1988 is also graphed. As demonstrated in Figure 9, the BLS telecommunications CPI estimates that since 1988, telecommunications prices have increased by 10.1 percent or an increase of 1.07

[54] See Diewert (1976).

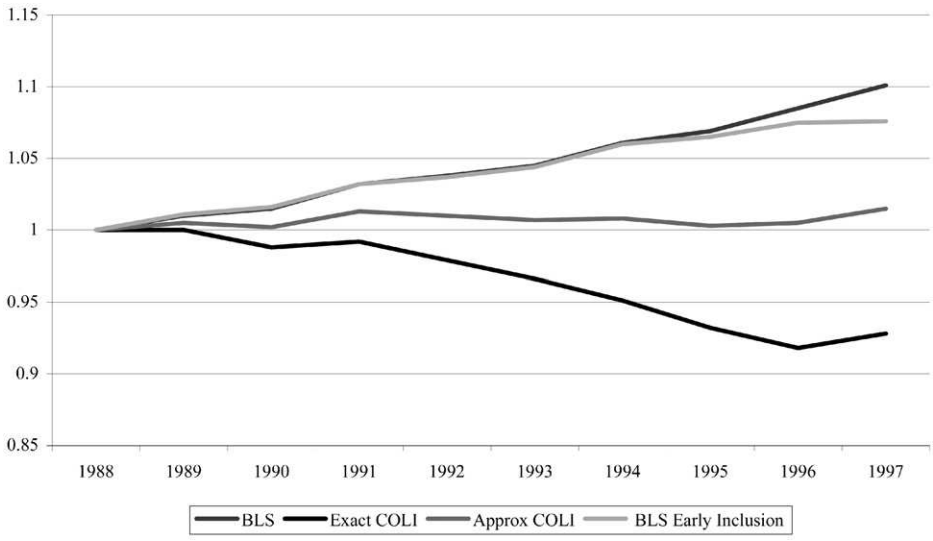


Fig. 9. Telecommunications CPI: BLS calculations and correct COLI calculations.

percent per year. This estimate ignores cellular service. A corrected telecommunication services COLI that includes cellular service decreased from 1.0 in 1988 to 0.928 in 1997 for a decrease of 0.8 percent per year. Next, I do similar calculations using the lower bound approximation to the COLI using equation (5). In Figure 9, this approximate telecommunication services CPI that includes cellular service is essentially unchanged since 1988 with the increase only 1.8 percent over the entire period[55].

5. Government regulation of mobile telephone

Governments have a long history of regulating telecommunications in North America and actually operating telecommunications networks in much of the rest of the world. The introduction of cellular technology in the mid-1980s coincided with the privatization and introduction of competition in fixed line telephone over the next decade[56]. Questions arose whether cellular should be regulated. Questions also arose over whether two or three providers were sufficient to provide competition and limit the exercise of market power. A further question was whether cellular was a 'luxury' service and need not be regulated, in contrast to landline telephone service, which was a 'basic' service that many countries desired

[55] Sensitivity tests for these estimates are found in Hausman (1999a).
[56] See the Chapter by Brock in this Handbook.

to be provided with universal service. In this section government regulation of cellular and its effects on consumers are reviewed[57]. First, the effects of the approximately ten-year long delay in the introduction of cellular telephone in the U.S. by the FCC are considered.

5.1. Estimation of the cost of regulatory delay in the U.S.

I now use the estimates from the previous section to estimate how much consumer welfare was lost by the original approximate ten year delay by the FCC in licensing cellular in the U.S[58]. Cellular technology was available for deployment in the U.S. by the early 1970s. Yet regulation by the FCC delayed the deployment of cellular in the U.S. by approximately ten years. The delay in provision of cellular telephone was caused by regulatory indecision and the subsequent licensing procedure used by the FCC, which was in charge of cellular spectrum. The FCC could not decide whether to allow AT&T to provide cellular service alone or to allow non-AT&T companies to provide cellular alone or to allow competition between the two groups. AT&T had invented cellular and argued because of significant economies of scale in spectrum usage that only one cellular provider should be present in each MSA[59]. Potential entrants into cellular argued that cellular could provide competition to AT&T's landline local monopoly at some time in the future so that AT&T should be barred from cellular.

Cellular telephone technology was sufficiently developed to begin operation in the early 1970s in the U.S., see Lee (1982) and Calhoun (1988) for histories of development of cellular telephone. In practice, cellular service in the U.S. did not begin until 1983. Clearly, the demand for mobile communications existed in the U.S. in the 1970's. Here I explain the delay caused by regulatory indecision and the subsequent licensing procedure used by the FCC.

The FCC made decisions and subsequently reversed itself. Finally, in the early 1980's the FCC decided to allow two cellular providers in each MSA, after cellular service had already begun in Scandinavia, Japan, and the U.K[60]. This duopoly situation was a competitive departure for the FCC (although competition did exist in the provision of IMTS, 'Improved Mobile Telephone Service,' which was the previous non-cellular generation car telephone service, as well as

[57] Government regulators unfortunately often attempt to protect competitors, rather than considering the interests of consumers. In the U.S. context and the failure of the FCC to promote consumer interests see Hausman and Shelanski (1999), and Hausman and Sidak (1999).

[58] This section is based on Hausman (1997).

[59] McKenzie and Small (1997) actually estimated decreasing returns to scale using U.S. data from 1993–1995.

[60] The FCC began its inquiry to reallocate additional spectrum for mobile telephone in 1968. By the time cellular telephone began operation in the U.S. in 1983, it had been in operation in both Scandinavia and Japan for over two years using the AMPS technology invented at Bell Labs.

paging service). The FCC decided to award 20 MHz of spectrum to each of the two cellular providers with 10 MHz of spectrum kept in reserve. In 1986 the FCC awarded 5 MHz of additional spectrum to the two cellular providers so that each now has 25 MHz of spectrum.

The FCC awarded the B block cellular frequency to the wireline telephone company in each MSA. Of course, this company was usually a BOC except for areas where GTE or an independent telephone company was awarded the spectrum. In a number of MSAs two or more wireline companies formed a partnership to operate the so-called 'wireline' network. To award the A block cellular frequency the FCC originally decided to conduct 'comparative hearings' (beauty contests) to decide who proposed the best cellular network. However, this procedure soon promised to create a morass of evidentiary and legal wrangling so that the FCC encouraged contenders to form partnerships. Companies such as Communications Industries, MCI, Metromedia, the Washington Post, and LIN Broadcasting became partnership members and were awarded the A block 'non-wireline' franchises. Because of procedural delays in awarding the non-wireline franchises, the B block networks typically began operation about 12–24 months earlier than the A Block non-wireline networks. Thus, a further delay was created. After realizing the fiasco of comparative hearings, the FCC subsequently used lotteries to award the non-wireline licenses in smaller MSAs and in RSAs. However, for these geographical areas it continued to award the B block license to the wireline carrier. Overall, the FCC delay was on the order of 7–10 years in the provision of cellular telephone in the U.S.

This regulatory indecision caused a new good, cellular telephone, to be unavailable in the U.S. when it was being offered in Scandinavia and Japan using technology invented by AT&T Bell Labs. This welfare loss is approximated by asking the question: if cellular had been available in 1983 with a 10 year history, but because of more limited and higher cost microprocessors and other semiconductor chips it cost twice as much (in 1983 dollars) as it did in 1994 and correspondingly, demand was lower because of the higher price, what was the lost consumer welfare?

The estimate of lost consumer welfare from the expenditure function of the previous section is approximately $49 billion in 1994 dollars. The lower bound approximation to the gain in consumer welfare from the introduction of cellular using the conservative linear approximation from the previous section is $24.2 billion. The gain in consumer welfare measured as the compensating variation from cellular is in the range of $24–49 billion per year, which demonstrates the substantial value to consumers from the introduction of cellular telephone.

My estimate of the loss in consumer welfare due to the ten-year delay, was approximately $24.3 billion in 1983 dollars or about $33.5 billion in 1994 dollars. Thus, the lost consumer welfare (compensating variation) was about

$76 per subscriber per month, which would be compared to an average monthly service price (with the 50 percent increase) of about $120 per month. Even if it is assumed that demand for cellular would only have been 1/2 as great in 1983 because of decreased functionality, there is still an estimated welfare loss of approximately $16.7 billion. This lost consumer welfare is extremely large and demonstrates how regulatory policy can significantly affect consumer welfare.

The consumer welfare cost of holding up the introduction of a new good is much larger than the effects of higher prices or other regulatory effects on demand, because the entire compensating variation is lost when regulatory delays cause demand to be zero. Looked at another way, the introduction of cellular has created significant value for consumers. Thus, new telecommunications services can improve consumer welfare by very large amounts. Regulatory delay can therefore have potentially large negative effects on the U.S. economy.

Again the possible question arises of why the FCC created such a large amount of harm to U.S. consumers and the U.S. economy. The FCC was confronted with a very difficult decision with respect to cellular. Delaying a difficult decision appeared to be the FCC's chosen response. Losses in consumer welfare arising from the regulatory delay did not appear to be involved in the FCC's regulatory approach. Indeed, if cellular service had not begun in other countries, which helped create pressure for the FCC to finally come to a decision, it is quite likely that the advent of cellular telephone service would have delayed for an even greater period in the U.S.

5.2. *Regulation of mobile prices*[61]

In the U.S. for the first 12 years of operation, cellular telephone operated as a duopoly. Given the significant fixed costs of cellular networks, the competitive outcome could not result in perfect competition[62]. In the U.S. each of 51 state regulatory commissions decided on whether to regulate cellular prices or to use market outcomes[63]. In an interesting natural experiment 26 states regulated cellular prices, while the other 25 did not. In Table 1, monthly service prices in 1994 are listed for the least expensive plan for average usage of 160 minutes per month (80 percent peak) for up to a 1-year contract.

The fact that regulation goes along with higher monthly service prices is evident from Table 1. Every regulated price in Table 1 is greater than every unregulated

[61] This section is based in part on Hausman (1995a, 1995b).

[62] For a discussion of possible non-competitive outcomes in the mobile industry, see Parker and Röller (1997).

[63] In the U.S. the District of Columbia acts as the 51st state.

Table 1
Average cellular prices in the top 10 MSAs: 1994 160 minutes of use (80 percent peak)[64]

MSA No.	MSA	Monthly price	Regulated
1	New York	$110.77	Yes
2	Los Angeles	$99.99	Yes
3	Chicago	$58.82	
4	Philadelphia	$80.98	
5	Detroit	$66.76	
6	Dallas	$59.78	
7	Boston	$82.16	Yes
8	Washington	$76.89	
9	San Francisco	$99.47	Yes
10	Houston	$80.33	

price in Table 1[65]. The average price of regulated MSAs is $98.10 while the average price of unregulated MSAs is $70.59, which is a difference of $27.51 per month or 39 percent.

Table 1 demonstrates clearly that regulation of cellular telephone led to higher prices for consumers in the U.S. Econometric analysis allows one to quantify the higher prices that consumers pay in regulated states. I have run a regression on cellular prices in the top 30 MSAs which accounts for MSA population, average commuting time, and average MSA income[66]. These top 30 MSAs contain about 107 million pops, or about 41 percent of the entire U.S. population. Regulation is treated as a jointly endogenous variable and uses instrumental variables in estimation. The coefficient of the regulation variable is 0.149, which means that regulated states have cellular prices that are 15 percent higher, holding other economic factors equal. The coefficient is estimated very precisely (standard error = 0.052) and the finding is highly statistically significant (t statistic = 2.87). Thus, states that regulate do have significantly higher cellular prices in large MSAs. Now in the top 30 MSAs overall, regulated prices are 23.6 percent higher. Thus, other economic factors explain about 9 percent of the higher prices and regulation explains 15 percent. Regulation is the major factor associated with the higher prices[67].

[64]This usage, 160 minutes per month, was the approximate average usage of cellular customers in 1994.
[65] The probability that every regulated price would exceed every unregulated price if the prices had no relationship to regulation is 0.00002.
[66] See Hausman (1995a, 1995b)
[67] See Duso (2000) for a recent paper, which considers the decision of states to regulate in terms of lobbying activity by U.S. cellular companies.

To explore this issue further I also collected data from cellular companies for the years 1989–93 and run a similar regression. Over this time period for 160 minutes of use regulation led to a higher price of 14.2 percent which is again estimated quite precisely (standard error = 0.029) and is very statistically significant (t statistic = 4.9). Thus, the results of the effect of regulation are very similar for the period 1989–1993 and for the single year 1994. I then did a similar econometric analysis, but used prices in the top 30 MSAs for 250 minutes per month for a typical 'high usage' customer. My estimate of the effect of regulation on price was 15.0 percent higher, which is again extremely statistically significant (t statistic = 6.57). Lastly, a similar econometric analysis was done for 30 minutes of monthly usage–a very light user of cellular. The coefficient estimate was 18.4 percent and it is highly statistically significant (t statistic = 4.2). Thus, econometric analysis demonstrated that in large MSAs for average usage of 160 minutes per month, for high usage of 250 minutes per month, and for low usage of 30 minutes per month that regulation is associated with higher prices of about 15 percent. Regulation led to higher prices, and it harmed consumers[68].

In 1993 Congress instructed the FCC to deregulate cellular prices unless a given state that was regulating cellular prices could show regulation was 'necessary.' Eight states petitioned the FCC to continue regulation, and the FCC turned them down in late 1994. One state appealed, but regulation completely ended in 1995. From Figure 5 above, it was previously noted that cellular prices decreased significantly in 1995–96 both because of new PCS entry and because of deregulation. As a test of econometric specification above, cellular (and PCS) prices in 1996 were taken and the previous specification reran. The regulation indicator variable was included to test the hypothesis whether it was regulation that caused prices to be higher in the regulated states, or whether it was some left out factor in the econometric model that would continue after prices were deregulated. The estimated coefficient on the (previously) regulated indicator variable is 0.9 percent (s.e. = 0.05). So regulation no longer had an effect in these states. Looked at another way, cellular prices fell after PCS entry, but they fell by about 14 percent more in states that had been previously regulated. I conclude that this excess 14 percent decrease was due to the elimination of regulation. Thus, when regulation was ongoing the prices in the regulated states (after controlling for other factors) were 15 percent higher than non-regulated states. In 1996 after price regulation was eliminated, the prices in the previously regulated states were not significantly different than the non-regulated states. Furthermore, the difference in prices was less than 1 percent. Thus, I conclude that state regulation of cellular prices led to higher prices to consumers by 14 to 15 percent. Regulators, by attempting to

[68] Ruiz (1995) did not find a similar effect of regulation using the prices of the average user. However, her results seem inconsistent with Table 1 and with the subsequent decrease in relative cellular prices once regulation was eliminated.

protect resellers and other competitors from competition, led to significantly higher prices to consumers.

5.3. *Regulation of mobile access and mandated roaming*

Mobile is now largely deregulated in the U.S. However, in the U.K. and in the European Union regulation to require access on mobile network providers to other competitors has been approved[69]. No pricing formulae have yet to be determined, but the process is expected to be quite contentious. However, the basic question should be why is forced access necessary? Unless the mobile companies can exercise significant market power, regulation is likely to harm consumers. By giving other competitors a free option for the use of the network providers' investments, the result is to discourage investment and innovative services[70].

Two recent regulatory decisions are extremely peculiar in this respect. In 1999 the U.K. regulatory agency, Oftel, required Cellnet and Vodafone to provide network access and mandated roaming, while not making a similar requirement of the two competitors Orange and One2One with respect to network access. Yet, Orange and One2One are significant competitors, with Orange's and One2One's net new customer additions during 1998 and 1999 the same or greater than Cellnet's. With no barriers to expansion for Orange and One2One and the regulatory finding that they lack market power, the requirement that Cellnet and Vodafone provide required access does not seem to make economic sense, except as a competitor protection measure. To the contrary, the Australia regulator, the ACCC, denied a request for mandated access in January 2000 and earlier denied a request for mandated roaming[71]. However, the ACCC takes as its primary regulatory goal the long-term interests of end users (LTIE), while Oftel the U.K. regulator does not have a regulatory goal of consumer welfare. Instead, it tends to put competitor welfare ahead of consumer welfare.

The recent decision in April 2000 by the European Commission to require Vodafone to provide network access to its competitors as a condition that Vodafone be permitted to purchase Mannesmann is even more curious. No increase in concentration resulted from the merger, and one of Vodafone's goals was to provide lower priced inter-European roaming to its customers, a strategy that has been very successful in the U.S. as I discussed above. Competitors

[69] I served as a consultant to Vodafone in the proceeding.

[70] By a free option I mean the ability to use the investment if it is successful, but not to help pay for the investment if it is not successful. For the distortionary effects caused by free options given to competitors see Hausman (1997, 1999b).

[71] No barriers to expansion exist given the ability of cell splitting discussed above and the economic factor that price significantly exceeds marginal (incremental) cost under competition because of the significant fixed costs in mobile networks.

claimed that they would be put at a disadvantage, and the European Commission required Vodafone to share its network with these competitors. In a non-telecommunications merger, this type of outcome would be almost unheard of. Consider the situation where Company A, an Internet switch manufacturer, buys Company B a fiber optics electronics manufacturer. No overlap in current products exist, but Company A announces it will be able to manufacture a better-combined product with Company B's technology. No competition regulatory agency would typically require Company A to allow another Company C access to Company B's technology. Otherwise, the adverse incentive to investment would limit future innovation and other companies would not invest since they could subsequently free ride off Company B's investment. However, regulatory agencies in charge of telecommunications seem to find it difficult to forbear from regulation, even when significant amounts of competition exist. One is put in mind of the famous statement, "It's no fun to be a regulator unless you get to regulate," (Anon).

5.4. *Regulation of wireline call termination on mobile*

In the U.S. up until the present, calls received on a mobile handset from a wireline origination have been paid for by the mobile customer[72]. In other countries a 'calling party pays' policy has been adopted where the wireline customer that originates the call is charged for it. As a consequence many U.S. mobile customers do not give out their mobile number with a higher proportion of originating calls in the U.S., about 80 percent, compared to other countries. However, the growing popularity of the 'bucket' plans in the U.S. has been decreasing this effect.

In a number of countries regulators and analysts have claimed that market power may be present for these fixed to mobile calls. Claims for regulation to lower termination prices have been made. In the U.K. fixed to mobile calls charges are claimed to be high, and Armstrong (1997) and in this Volume states that high call termination charges may be used to cross-subsidize new mobile subscribers. Doyle and Smith (1998) find in a model of competition that regulation is not required, but that the US receiver party pays approach would solve the problem. Recently, the Australian regulator, the ACCC, has examined the issue of mobile termination charges. Gans and King (1999) claim that a mobile operator has an 'effective monopoly' over its customers and that regulation is needed because of wireline consumer ignorance over the terminating charges when they place a call[73]. Gans and King (1999) call for terminating charges to be set at marginal cost by the regulator. However, Wright (2000) demonstrates that above cost termination charges lead to lower cellular price and high penetration rates because of competition among mobile companies.

[72] The FCC has been considering a change in this policy to calling party pays.
[73] I consulted for Optus in this proceeding. See Hausman (2000b).

Almost all participants in the debate acknowledge that competition among mobile providers works well, with increasing entry and decreasing prices. Thus, no regulation is needed here. The question is who will pay for the fixed and common costs of the mobile network; mobile subscribers or fixed to mobile callers? The Gans-King proposal of setting terminating charges at marginal cost seem to make little economic sense. As Hausman (2000b) demonstrates, the standard Ramsey problem of the efficient method to cover the fixed and common costs will have both sets of customers paying above marginal cost to cover the fixed and common costs of the mobile networks. Indeed, given the estimated elasticities, the terminating call customers would pay a higher markup in an optimal solution than would the mobile originating call customers. Of course, a two-part tariff arrangement could be explored, but it is doubtful that regulators would require companies to pay a fixed charge for their customer's calls.

If a problem exists, it is because of customer ignorance of the charges they will pay for their terminating call. However, whether customer ignorance is a problem is not clear since many calls are repeat calls and customers receive itemized bills for their calls. But price regulation is not required to solve the problem if customer ignorance causes the problem. For terminating calls on mobile the operator could be required to identify itself, just as AT&T has done for many years in the U.S. when a long distance call is made on its wireline network. Consumer information would solve the potential market failure problem without the need for regulatory interference in competition and market determined prices. Both regulators and economists should first determine if a problem exists and then seek to solve the source of the problem, rather than turning to the highly distortionary solution of setting regulated prices, which has not worked well in the past in similar situations.

6. Taxation of mobile services[74]

In the U.S. federal, state, and local government authorities are now levying a wide range of taxes and fees on the use of mobile telephone services. The total effect of the FCC-imposed fees and other federal taxes is currently 4.52 percent (including the federal excise tax and the FCC's share of current universal service program funding)[75]. State and local taxes on wireless vary by jurisdiction, and commonly impose higher tax rates on wireless than on other businesses. Many localities charge a variety of direct and indirect fees to wireless providers. For example, the aggregate New York state tax is 20 percent so that the combined federal and state tax burden on a wireless user in New York is 24.5 percent or approximately $170 per year. Similar taxes on wireless users in California and Florida average about 21 percent. The resulting state tax obligation for the average wireless customer in

[74] This section is based on Hausman (2000a).

[75] See Hausman (1998) and Hausman and Shelanski (1999) for a discussion of the FCC taxation program to provide Internet subsidies for schools and libraries.

these states exceeds \$152 per year[76]. When federal taxes are included, the overall tax rate increases to 25.5 percent, and average consumers' tax bills increase to \$185 per year. Even in states where lower taxes are levied on wireless, the median state tax rate is 10 percent, and the tax payment for the average wireless customer is about \$62. Including current federal taxes, the median tax rate is 14.5 percent and the yearly tax bill is about \$91.

Economists have well-developed tools, used for more than 100 years, to assess the distortionary effect of a tax on consumer welfare and economic efficiency. Taxes decrease the consumption of a good or service and, in this case, also lead to under-utilization of the infrastructure investment made by wireless providers. In this section, the effect on deadweight loss and economic efficiency from the distortionary effect of taxation on wireless are calculated. The change in economic efficiency accounts for both harm to consumers, from the deadweight loss term, and harm to wireless providers who have invested tens of billions of dollars in their networks and who have paid over ten billion dollars to acquire their licenses, often from the federal government[77].

6.1. Estimation of economic efficiency losses

Taxes (and subsidies) distort economic activity. Taxes increase prices and thus lead to lower demand. This lower demand has two adverse affects on economic efficiency which is defined (approximately) as the sum of producer surplus and consumer surplus[78]. To the extent that the industry is imperfectly competitive and price exceeds marginal cost to cover fixed costs, decreased demand reduces the amount of producer surplus, which is the product of quantity demanded times the difference between price and marginal cost. Decreased demand from higher prices also affects consumers adversely since consumer surplus decreases. Thus, the change in economic efficiency from the imposition of a tax is given approximately by the formula:

$$\begin{aligned}\Delta E &\approx [-\Delta q_i(p_i - m_i) - .5\Delta p_i \Delta q_i] \\ &\approx \left[\eta_i \frac{\Delta p_i}{p_i}(p_i q_i - m_i q_i) + .5\eta_i \left(\frac{\Delta p_i}{p_i}\right)^2 p_i q_i\right] \end{aligned} \tag{8}$$

[76] For instance, a wireless user in California pays the following taxes: FCC taxes for the high cost fund, universal service and school and library internet subsidy, state, county and city sales taxes, taxes (fees) levied by the California Public Utilities Commission for universal service (3.2 percent), emergency telephone service (0.72 percent), high cost funds (3.14 percent), teleconnect fund (0.41 percent), hearing impaired fund (0.36 percent), local utility taxes (7 percent), and the federal excise tax (3 percent).

[77] PCS providers bought their spectrum in auctions conducted by the federal government. The cellular spectrum and much of the ESMR spectrum was distributed free of charge by the federal government in an earlier period before auctions were used. But, in many instances, current licensees obtained their spectrum by buying it at market prices from the original licensees.

[78] See, e.g. Auerbach (1985) for a further discussion of how taxation creates efficiency losses to the economy.

where the first term is the change in producer surplus and the second term is the change in consumer surplus, after the amount raised by the tax is subtracted[79]. Figure 10 provides a graphical demonstration of this relationship. Equation (8) demonstrates that taxes which cause prices to increase create losses in economic efficiency with the size of the efficiency loss depending on the price elasticity η_i, the magnitude of the price increase ($\Delta p_i/p_i$), the revenue of the good or service being taxed $p_i q_i$, and the marginal cost of production, m_i[80]. For mobile telephone the price elasticity η_i has been estimated to be -0.51, Hausman (1997), which is relatively high for telecommunications services[81]. The magnitude of the price increase ($\Delta p_i/p_i$) depends on the tax rate in each state with the median tax rate 10 percent and the high tax rates in the range of 20 percent-21 percent in California, New York, and Florida[82]. When current federal taxes are included, the median tax rate is 14.5 percent and in the high tax states the overall rate is 24.2 percent to 25.5 percent. The revenue of wireless service $p_i q_i$ is about \$525 per year, excluding taxes. Lastly the marginal cost of production for wireless, m_i, is relatively low, which is expected given the large fixed costs of wireless networks. The marginal cost is estimated to be about \$0.05 per minute[83]. Thus, the expected result from Equation (8) is relatively high efficiency costs to wireless taxation given the relatively high demand elasticity, the significant tax rates, and the low marginal cost of production.

Using equation (8), the cellular elasticity estimate considered above, and the fact that the marginal cost of wireless is about \$0.05 per minute while the median tax rate is 14.5 percent[84], I estimate that for average revenue raised by the tax on wireless:

$$\Delta E = 0.534^{*} TR \tag{9}$$

[79] Thus, as discussed above, the possible distortion created by expenditure of the tax is not considered. All the quantities in the formulae are assumed to be Hicksian compensated quantities. See Hausman (1981) for computation of compensated quantities.

[80] A more accurate estimation approach using the technique of Hausman (1981) is discussed in Hausman (2000a).

[81] However, Nattermann (1999) estimates the industry price elasticity in Germany to be -2.3.

[82] In these calculations I put sales taxes in the 'numeraire' price of other goods under the assumption that it is paid on purchases of other goods and services.

[83] To estimate this marginal cost I take the expected growth rate of a given cellular company and perturb the growth rate up or down by a small amount. I take the difference in the present value of costs and divide it by the levelized increase in output. Thus, the estimate takes account of more rapid expansion (e.g. cell splitting) by the cellular company as well an increase in its variable costs which also include payment to local exchange companies and to interexchange carriers for the increased traffic. The main economic fact that price greatly exceeds marginal cost is not significantly affected by the exact details of the calculation. This relationship between price and marginal costs holds for almost all telecommunications services.

[84] This estimate of marginal cost includes customer acquisition cost which is amortized over three years, which follows from an estimate of the average churn rate for new cellular customers.

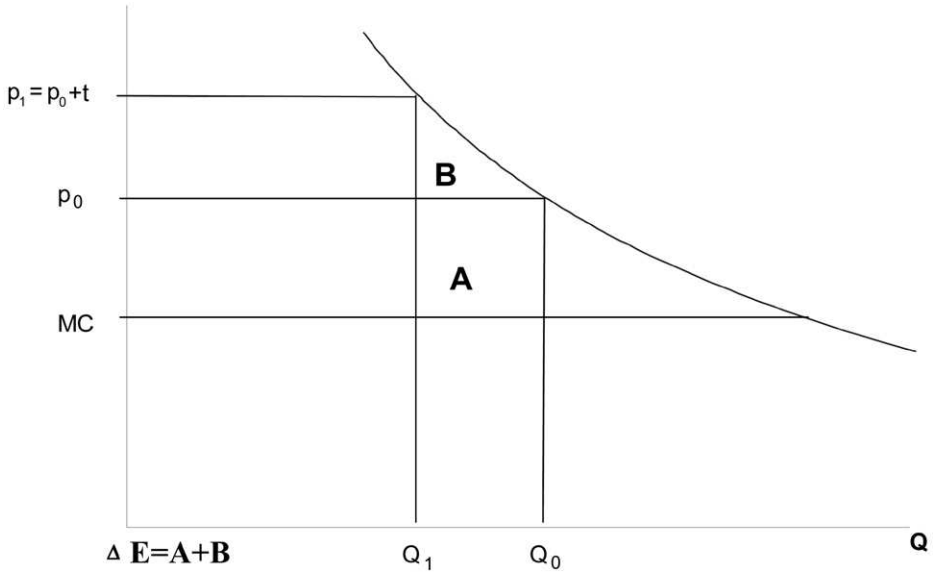


Fig. 10. Efficiency effect of tax.

where TR is total tax revenue raised. This calculation follows from dividing through Equation (8) by the tax revenue raised, $t_i p_i q_i$ (i.e. the tax revenue term TR), and using wireless revenue and tax amounts collected from wireless by the federal, state, and local jurisdictions.

For the high tax states the efficiency loss increases on average for each dollar of tax revenue raised from wireless customers. For a 21 percent state tax rate used in California, Florida and New York, the estimated efficiency loss increases to approximately \$0.70 for each dollar of tax revenue raised. For lower tax rates, the estimated efficiency loss decreases accordingly. The elasticity estimate that was used to calculate equation (9), reported in Hausman (1997), is −0.51. This estimate was based on data up through 1993. Using more recent data, it has been estimated as an elasticity of about −0.71, although the estimate is not statistically significantly different from the earlier estimate[85]. An increased elasticity estimate might be expected given the rapid penetration of mobile telephone and the expected results that early adopters place a higher marginal valuation on their usage while later adopters are affected by decreasing prices. Thus, the higher penetration among residential users may be leading to a higher price elasticity,

[85] Recent estimates of the price elasticity in Australia have estimated an elasticity of approximately −0.8, which is close to my estimates.

since these non-business users have a higher price elasticity[86]. While one would need to collect more data before changing the elasticity estimate, note from equation (8) that the estimated efficiency loss is homogeneous of degree one in the elasticity estimate. Thus if the higher elasticity estimate were used to estimate the average efficiency loss, the amount would increase from 0.534 for the median tax state to 0.743. The efficiency loss for the high tax states would also increase accordingly.

Perhaps a more relevant calculation is the marginal efficiency loss to the economy from changes to the tax rates. The FCC and state and local tax authorities apparently view wireless as a ready tax revenue source so that they have been increasing the tax rates over time. The FCC has increased tax rates to provide universal service for landline telephone users and to provide Internet subsidies to schools and libraries[87]. These subsidies and the taxes (fees) to fund them are expected to continue increasing in the near future. The formula for the marginal efficiency loss of increased taxation is computed by taking the marginal change in equation (8) with respect to the tax rate, $\partial \Delta E/\partial t_i$, and dividing by the marginal change in tax revenue with respect to the tax rate, $\partial TR/\partial t_i$:

$$\left(\frac{\partial \Delta \mathrm{E}/\partial t_i}{\partial TR/\partial t_i}\right) \approx \frac{\eta_i\left(1-\frac{m_i}{p_i}\right)+\eta_i\frac{t_i}{p_i}+\left[\eta_i\frac{t_i m_i}{p_i^2}-.5\eta_i\frac{t_i^2}{p_i^2}\right]\frac{\partial p_i}{\partial t_i}}{1-\eta_i\frac{t_i}{p_i}\frac{\partial p_i}{\partial t_i}} \qquad (10).$$

Using equation (10) together with the assumption that $\partial p_i/\partial t_i = 1$ along with the fact that $t_i/p_i = 0.1452$ for the median state when federal taxation is included, Equation (10) is estimated to be 0.709[88]. When the marginal efficiency loss is

[86] Under various competitive assumptions that businesses pass along the increased cost of inputs in their final product prices, the estimates in this paper would remain approximately the same. However, as I discuss subsequently, the pass through assumptions can be affected by oligopoly models of final product demand. So as not to unduly complicate the models, I assume unitary pass through of the wireless taxation to final product demand and assume that the same price elasticities hold for business and residential customers. Thus, business demand can be interpreted as a derived demand for wireless that arises from final product demand from consumers.

[87] Hausman (1998) and Hausman and Shelanski (1999) discuss the Internet subsidy program for schools and libraries.

[88] If instead of the assumption that $(\partial p_i/\partial t_i = 1)$, I use a differentiated product oligopoly markup model assumption along with constant elasticity demand curves, the marginal efficiency loss could be higher than 0.71. Other oligopoly models, especially models based on linear demand curves could find $(\partial p_i/\partial t_i < 1)$. For a further discussion of these matters see e.g. Hausman and Leonard (1999). However, the introduction of 4–5 new wireless competitors in addition to the two cellular incumbents in each market will lead to increased competition, leading to the conclusion that the entire tax will be passed on to customers. When I did a similar analysis for the introduction of the 'E-rate' tax on long distance in Hausman (1998), the chief economist at the FCC at that time claimed that the long distance companies might not pass on the increased tax to their customers. However, experience demonstrates that the long distance companies did pass on the entire tax, often using a separate line item on the bill to call customer attention to the tax.

calculated using the exact calculation based on Equation (9) using the approach of equation (10) instead of the traditional approximation, the marginal efficiency loss is estimated to be 0.724. For all further calculations, the exact approach is used rather than the approximation of Equation (10).

Thus, for increased taxation the efficiency loss to the economy is approximately \$0.72 for each \$1 of additional tax revenue. For states such as California and New York with tax rates near 20 percent (which rise to 25 percent when current federal taxation is included), the marginal efficiency loss is about \$0.93 for each \$1 of tax revenue raised from wireless service. Thus, the marginal efficiency loss is quite high since for each dollar raised by an increase in wireless taxes, \$0.72 to \$0.93 of efficiency loss is created for the economy, beyond the tax revenue raised[89].

Three reasons exist for this high amount of efficiency loss to the economy from wireless taxation, which can be seen by an examination of equations (8) and (10): (1) the elasticity η_i is relatively high, (2) m_i/p_i is relatively low since gross margins are high in wireless which is to be expected given the large fixed costs of wireless networks, and (3) t_i/p_i is relatively high in the range of 14 percent-25 percent.

6.2. Comparison with alternative taxes

Rather than taxing telecommunications usage to fund the subsidy for universal service landline telephone subsidies and Internet access subsidies for schools and libraries, Congress could have used general tax revenue. Similarly, states can levy taxes on incomes or expenditures (sales taxes) to fund their various social programs. While no generally agreed to number exists for the value of the marginal efficiency loss to the economy from increasing overall taxes, the range of estimates is reasonably close. The generally used estimates are all below about \$0.40 of marginal efficiency loss per dollar of additional revenue raised[90]. Thus, they are all significantly less than the \$0.72 to \$1.14 efficiency loss per additional dollar of tax revenue raised by the FCC and by the state and local authorities. Thus, considerably less expensive means to raise tax revenues, in terms of economic efficiency losses, exist for the federal government and for the states in terms of increasing income taxes or other broad-based taxes, rather than targeting the use of wireless services. Government taxation of mobile telephone imposes an especially high efficiency loss on the economy given the demand and cost factors that exist for mobile telephone. Governments would be much better off to use a broad based tax such as an income tax or find alternative goods to tax with lower demand elasticities of a lower proportion of fixed costs to total costs.

[89] Since the numerator is homogeneous of degree one in the price elasticity and the denominator will decrease with a higher elasticity, the marginal efficiency loss would increase significantly if the higher elasticity of -0.71, which I discussed above, were used. Indeed the estimate of \$0.724 increases to \$1.14 for the marginal dollar of tax revenue.

[90] See Hausman (2000a) for the sources of these estimates.

In this section the outcome of three government regulatory and taxation policies for mobile telephone has been considered. All three policies had a significant and large negative effect on consumer welfare. While mobile telephone is a telecommunications service from a sector of the economy that has traditionally been heavily regulated, a clear lesson is believed to be that regulation should be removed from mobile telephone[91]. Regulation of a dynamic technology is notoriously difficult, and U.S. regulation of mobile telephone has provided a very good example. Sufficient competition now exists in the U.S. and most other countries so that regulation of mobile telephone is unnecessary. However, regulators find it notoriously difficult to 'let go' since they seemingly have little faith in the operations of markets. In the U.S. it took an act of Congress to deregulate mobile. While economics students learn that in the absence of market failure consumers benefit from market outcomes, regulators either favor competitors' interests over consumer welfare because of political considerations or do not believe in economics. One leaves it to the reader to decide.

7. Conclusion

Successful new telecommunications services create very large gains in consumer welfare. Mobile telephone is an example of an extremely successful new telecommunication service that already has changed how people live and work. In any large urban area, it is difficult to walk more than a block or two without seeing someone using a mobile telephone. Mobile telephone penetration has surpassed landline telephone penetration in Scandinavia and Japan, and it will do so in a number of other European and North American countries in the next few years. Thus, the initial technology invented by Bell Labs and advanced by numerous companies since, has led to an innovative product of great usefulness. With the possible future combination of wireless telephones and PDAs with the Internet, the usefulness may increase by even more. Indeed, the future of landline telephone may be called into question by the availability of mobile telephones.

Government decisions have played a significant role in the past and future of mobile telephones. Government decisions did permit competition in mobile, in contrast to the single provider model of landline telephone. However, at least in the U.S., the FCC delayed the introduction of cellular telephones, costing consumers tens or even hundreds of billions of dollars in lost consumer welfare. Again in the U.S., it was the Congress that forced the regulatory agencies to auction off spectrum rather than use beauty contests, and used political influence to allocate scarce spectrum. Further, Congress caused cellular prices to be deregulated in the U.S. Thus, the lesson that one draws from the experience is that consumers and the mobile industry, in terms of lower prices and greater

[91] Regulation may be needed to minimize spectrum interference, but economic regulation is unnecessary and generally harmful in this situation.

investment, would have been better off, and would be better off in the future, with de-regulation of mobile telephone. Competition among 4 to 5 providers should be sufficient to protect consumers. In many industries with 4 to 5 providers that arise because of large fixed and sunk costs, no responsible economist would call for price or access regulation. Yet, telecommunications regulators, perhaps because of their experience in regulation of landline telephone, seem to continue to regulate an industry that does not need regulation. Competition is typically superior to regulation in situations similar to the mobile industry, and regulators seem to want to engage in competitor protection policies. Instead, regulators should consider consumer welfare and allow competition to take care of consumers.

References

Ahn, H. and M. Lee, 1999, An econometric analysis of the demand for access to mobile telephone networks, Information Economics and Policy, 11, 3, 297–305.

Armstrong, M., 2002, The theory of access pricing and interconnection, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science B. V., 295–384.

Armstrong, M., 1997, Mobile telephony in the UK, Oxford mimeo.

Barros, P. and N. Cadima, 2000, The impact of mobile phone diffusion on the fixed-link network, Mimeo CEPR.

Brock, G., 2002, Historical overview, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science B. V., 43–74.

Calhoun, G., 1998, Digital cellular radio, Norwood, MA: Artech.

Cramton, P., 2002, Spectrum auctions, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science BV, 605–640.

Crandall, R. and J. Hausman, 2000, Competition in U.S. telecommunications services four years after the 1996 Act, in S. Peltzman and C. Winston, eds., Deregulation of Network Industries, Washington DC: Brookings Institution.

Diewert, W. E., 1976, Exact and superlative index numbers, Journal of Econometrics, 4, 2, 115–45.

Doyle, C. and J. Smith, 1998, Market structure in mobile telecoms, Information Economics and Policy, 10, 4, 471–88.

Duso, T., 2000, Who decides to regulate? Lobbying activity in the U.S. cellular industry, Mimeo.

Gans, J. and S. King, 1999, Termination charges for mobile networks, Mimeo, Melbourne University.

Gruber, H. and F. Verboven, 2000, The diffusion of mobile telecommunications services under entry and standards regulation, Mimeo, London: CEPR.

Hausman, J., 1979, Individual discount rates and the purchase and utilization of energy using durables, Bell Journal of Economics, 10, 1, 33–54.

Hausman, J., 1981, Exact consumer's surplus and deadweight loss, American Economic Review, 71, 4, 662–676.

Hausman, J.,1995a, The cost of cellular telephone regulation, Working Paper, MIT.

Hausman, J., 1995b, State regulation of cellular prices, Wireless Communications Forum, 3, 61–68.

Hausman, J., 1997, Valuation of new goods under perfect and imperfect competition, in T. Bresnahan and R. Gordon, eds., The Economics of New Goods, Chicago: University of Chicago.

Hausman, J., 1997, Valuing the effect of regulation on new services in telecommunications, Brookings Papers on Economic Activity, Microeconomics, 1–38.

Hausman, J., 1998, Taxation by telecommunications regulation, Tax Policy and the Economy, 12, 29–48.

Hausman, J., 1999, Cellular telephone, new products and the CPI, Journal of Business and Economics Statistics, 17, 188–194.
Hausman, J., 1999b, The effect of sunk costs in telecommunication regulation, in J. Alleman and E. Noam, eds, The New Investment Theory of Real Options and its Implications for Telecommunications Economics, Boston: Kluwer, 191–204.
Hausman, J., 2000a, Efficiency effects on the U.S. economy from wireless taxation, National Tax Journal, 53, 733–42.
Hausman, J., 2000b, Declaration of Prof. J. Hausman to the ACCC.
Hausman, J. and G. Leonard, 1999, Efficiencies from the consumer viewpoint, George Mason Law Review, 7, 707–727.
Hausman, J. and W. Newey, 1995, Non-parametric estimation of exact consumer surplus and deadweight loss, Econometrica, 63, 1445–1476.
Hausman, J. and H. Shelanski, 1999, Economic welfare and telecommunications regulation: The E-rate policy for universal service subsidies, Yale Journal On Regulation, 16, 19–51.
Hausman, J. and J. G. Sidak, 1999, A consumer-welfare approach to the mandatory unbundling of telecommunications networks, Yale Law Journal, 109, 417–505.
Hicks, J. R., The valuation of the social income, Economica, 7, 105–124.
Jha, R. and S. K. Majumdar, 1999, A matter of connections: OECD telecommunications sector productivity and the role of cellular technology diffusion, Information Economics and Policy, 11, 243–269.
Lee, W., 1982, Mobile communications engineering, New York: McGraw-Hill.
McKenzie, D. and J. Small, 1997, Econometric cost structure estimates for cellular telephony in the U.S., Journal of Regulatory Economics, 12, 147–157.
Nattermann, P., 1999, The German cellular market, Info, 1, 355–365.
Parker, P. M. and L. H. Röller, 1997, Collusive conduct in duopolies: Multimarket contact and cross-ownership in the mobile telephone industry, RAND Journal of Economics, 28, 304–322.
Rothbarth, E., 1941, The measurement of changes in real income under conditions of rationing, Review of Economic Studies, 8, 100–107.
Ruiz, L., 1995, Pricing strategies and regulatory effects in the U.S. cellular telecommunications duopolies, in G. Brock, ed., Towards a Competitive Telecommunications Industry: Selected Papers from the 1994 Telecommunications Policy Research Conference, Mahwah, NJ: Lawrence Erlbaum.
Valletti, T., 1999, A model of competition in mobile communications, Information Economics and Policy, 11, 61–72.
Valletti, T., 1999b, Switching costs in vertically related markets, Mimeo, London School of Economics.
Valletti, T. and M. Cave, 1998, Competition in UK mobile communications, Telecommunications Policy, 22, 109–131.
Wright, J., 2000, Competition and termination in cellular networks, Mimeo, University of Auckland.

Chapter 14

SPECTRUM AUCTIONS*

PETER CRAMTON*

University of Maryland

Contents

* I am grateful to the National Science Foundation for funding. I advised several governments on auction design and several bidders on auction strategy. The views expressed are my own.

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

From July 1994 to February 2001, the Federal Communications Commission (FCC) conducted 33 spectrum auctions, raising over $40 billion for the U.S. Treasury. The auctions assigned thousands of licenses to hundreds of firms. These firms are now in the process of creating the next generation of wireless communication services. The FCC is not alone. Countries throughout the world now are using auctions to assign spectrum. Indeed, the early auctions in Europe for third-generation (3G) mobile wireless licenses raised nearly $100 billion. Auctions have become the preferred method of assigning spectrum.

The FCC auctions have shown that using an auction to allocate scarce resources is far superior to the prior methods: comparative hearings and lotteries. With a well-designed auction, there is a strong tendency for the licenses to go to the parties that value them the most, and the Treasury obtains much-needed revenues in the process.

Overall, the auctions have been a tremendous success, putting essential spectrum in the hands of those best able to use it. The auctions have fostered innovation and competition in wireless communication services. Taxpayers, companies, and especially consumers have benefited from the auctions. In comparison with other countries, the FCC auctions represent the state-of-the-art in spectrum auction design and implementation. The FCC began its auctions with an innovative design, and has continued to improve the auctions since then. The FCC's leadership in spectrum auctions has had positive consequences world-wide. Many countries wisely have imitated the FCC auctions; those that have not have suffered from inefficient license assignments and other flaws.

All but two of the FCC auctions have used a simultaneous ascending design in which groups of related licenses are auctioned simultaneously over many rounds of bidding. In each round, bidders submit new higher bids on any of the licenses they desire, bumping the standing high bidder. The auction ends when a round passes without any bidding; that is, no bidder is willing to raise the price on any license. This design is a natural extension of the English auction to multiple related goods. Its advantage over a sequence of English auctions is that it gives the bidders more flexibility in moving among licenses as prices change. As one license gets bid up, a bidder can shift to an alternative that represents a better value. In this way, bidders are able to arbitrage across substitutable licenses. Moreover, they can build packages of complementary licenses using the information revealed in the process of bidding.

There is now substantial evidence that this auction design has been successful. Revenues often have exceeded industry and government estimates. The simultaneous ascending auction may be partially responsible for the large revenues. By revealing information in the auction process, bidder uncertainty is reduced, and the bidders safely can bid more aggressively. Also, revenues may increase to the

extent the design enables bidders to piece together more efficient packages of licenses.

Despite the general success, the FCC auctions have experienced a few problems from which one can draw important lessons. One basic problem is the simultaneous ascending auction's vulnerability to revenue-reducing strategies in situations where competition is weak. Bidders have an incentive to reduce their demands in order to keep prices low, and to use bid signalling strategies to co-ordinate on a split of the licenses. I identify problems with the FCC and other spectrum auctions, and discuss how to minimise these problems.

I begin by explaining why it makes sense to auction the spectrum. Then I discuss auction design. The simultaneous ascending auction is discussed in detail, and its performance is evaluated. I describe how the auction format has evolved in response to perceived problems. The FCC's recently proposed approach to package bidding is discussed. I examine the UMTS auctions in Europe. Finally, I provide advice to governments engaged in spectrum auctions.

2. Why auction the spectrum?

There is substantial agreement among economists that auctions are the best way to assign scarce spectrum resources (McMillan, 1995). Auctions ask an answer to the basic question 'Who should get the licenses and at what prices?' Ronald Coase (1959) proposed auctioning spectrum over forty years ago, yet it was not until 1994 that spectrum auctions became a reality in the US. Hazlett (1998) provides an analysis of the history that ultimately led to spectrum auctions.

Alternatives to auctions are an administrative process or lotteries. Both alternatives were used and then rejected in the US.

An administrative process has those that are interested in the spectrum make a proposal for how they intend to use it. This approach commonly is referred to as a 'beauty contest.' After hearing all the proposals, the regulator awards spectrum to those with the most attractive proposals. Beauty contests suffer from several problems. First, they are extremely slow and wasteful. Even with streamlined hearings, it took the FCC an average of two years to award thirty cellular licenses. Competitors spend vast sums trying to influence the regulator's decision. Second, beauty contests lack transparency. It is difficult to see why one proposal won out over another. Worse yet, the ability of the regulator to successfully identify the best proposals is limited.

Unacceptable delays in assigning cellular licenses led the FCC to switch to lotteries. With a lottery, the FCC randomly selects license winners from among those that apply. The problem here is that since the licenses are enormously valuable there is a strong incentive for large numbers to apply. Indeed, the FCC received over four hundred thousand applications for its cellular lotteries.

The huge number of applications wasted resources in creating and processing the applications. Moreover, the winners were not those best suited to provide a service. It took years for the licenses to be transferred via private market transactions to those capable of building out a service. Lotteries were quickly abandoned in favour of auctions.

The primary advantage of an auction is its tendency to assign the spectrum to those best able to use it. This is accomplished by competition among license applicants. Those companies with the highest value for the spectrum likely are willing to bid higher than the others, and hence tend to win the licenses. There are several subtleties, which are addressed below that limit the efficiency of spectrum auctions. Still a well-designed auction is apt to be highly efficient. A second important advantage of auctions is that the competition is not wasteful. The competition leads to auction revenues, which can be used to offset distortionary taxation. Finally, an auction is a transparent means of assigning licenses. All parties can see who won the auction and why.

Despite their virtues, standard auctions at best ensure that the bidder with the highest private value wins, rather than the bidder with the highest social value. Private and social values can diverge in these auctions because the winners will be competing in a marketplace. One collection of winners may lead to a more collusive industry structure. For example, a license may be worth more to an incumbent than a new entrant, simply because of the greater market power the incumbent would enjoy without the new entrant. Recognising this, the regulator typically limits the amount of spectrum any one firm can hold in any geographic area.

3. Auction design

Spectrum auctions typically involve the sale of multiple interrelated licenses. We have only begun to understand such auctions in theory and practice. We neither have strong theoretical nor empirical results to guide the design of spectrum auctions. Designing spectrum auctions is as much art as it is science.

The objective of most spectrum auctions is two-fold. The primary goal is efficiency – getting the spectrum in the hands of those best able to use it. A secondary goal is revenue maximisation. Even efficiency-minded governments should care about the revenues raised at auction, since auction revenues are less distortionary than the principal source of government revenues – taxation. Economists estimate that the welfare loss from increasing taxes in the United States is in the range of 17 – 56 cents per dollar of extra revenue raised (Ballard et al., 1985). Hence, in designing the auction, the government should care about revenues. Another common objective is increasing competition for wireless services.

Sometimes it is argued that efficiency is not an important objective. If resale is allowed, will not postauction transactions fix any assignment inefficiencies?

The answer is "yes" in a Coasean world without transaction costs. However, transaction costs are not zero. Postauction transactions often are made difficult by strategic behaviour between parties with private information and market power. Efficient auctions are possible before assignments are made but may become impossible after an initial assignment. The problem is that the license holder exercises its substantial market power in the resale of the license (Cramton et al., 1987). For this reason, it is important to get the assignment right the first time. Moreover, efficient auctions tend to raise substantial revenues, especially when resale is possible (Ausubel and Cramton, 1999).

Much of the research on spectrum auction design has focused on the Personal Communication Services (PCS) auctions in the US. Below I summarise several of the important issues, and the FCC's ultimate conclusion. These issues are discussed in several papers (e.g., Cramton, 1997; McMillan, 1994; Milgrom, 2000; and Chakravorti et al., 1995; Krishna and Rosenthal, 1997; Rosenthal and Wang, 1997).

3.1. Open bidding is better than a single sealed bid

An essential advantage of open bidding is that the bidding process reveals information about valuations. This information promotes the efficient assignment of licenses, since bidders can condition their bids on more information. Moreover, to the extent that bidder values are affiliated, it may raise auction revenues (Milgrom and Weber, 1982), since the winner's curse is reduced. Bidders are able to bid more aggressively in an open auction, since they have better information about the item's value.

The advantage of a sealed-bid design is that it is less susceptible to collusion (Milgrom, 1987). Open bidding allows bidders to signal through their bids and establish tacit agreements. With open bidding, these tacit agreements can be enforced, since a bidder can immediately punish another that has deviated from the collusive agreement. Signalling and punishments are not possible with a single sealed bid.

A second advantage of sealed bidding is that it may yield higher revenues when there are *ex ante* differences among the bidders (Maskin and Riley, 1995; Klemperer, 1998). This is especially the case if the bidders are risk averse (Maskin and Riley, 1984; Matthews, 1983). In a sealed bid auction, a strong bidder can guarantee victory only by placing a very high bid. In an open auction, the strong bidder never needs to bid higher than the second-highest value.

In the PCS spectrum auctions, there was a consensus among experts in favour of open bidding. The advantage of revealing more information in the bidding process was thought to outweigh any increased risk of collusion. Collusion was viewed as unlikely, and revenue maximisation was a secondary goal.

3.2. Simultaneous open bidding is better than sequential auctions

A frequent source of debate was whether licenses should be sold in sequence or simultaneously. A disadvantage of sequential auctions is that they limit information available to bidders and limit how the bidders can respond to information. With sequential auctions, bidders must guess what prices will be in future auctions when determining bids in the current auction. Incorrect guesses may result in an inefficient assignment when license values are interdependent. A sequential auction also eliminates many strategies. A bidder cannot switch back to an earlier license if prices go too high in a later auction. Bidders are likely to regret having purchased early at high prices, or not having purchased early at low prices. The guesswork about future auction outcomes makes strategies in sequential auctions complex, and the outcomes less efficient.

In a simultaneous auction, a large collection of related licenses is up for auction at the same time. Hence, the bidders get information about prices on all the licenses as the auction proceeds. Bidders can switch among licenses based on this information. Hence, there is less of a need to anticipate where prices are likely to go. Moreover, the auction generates market prices. Similar items sell for similar prices. Bidders do not regret having bought too early or too late.

The Swiss wireless-local-loop auction conducted in March 2000 illustrates the difficulties of sequential sale. Three nationwide licenses were sold in a sequence of ascending auctions. The first two licenses were for a 28 MHz block; the third was twice as big (56 MHz). Interestingly, the first license sold for 121 million francs, the second for 134 million francs, and the third (the large license) sold for 55 million francs. The largest license sold for just a fraction of the prices of the earlier licenses.

Proponents of sequential auctions argue that the relevant information for the bidders is the final prices and assignments. They argue that simultaneous auctions do not reveal final outcomes until the auction is over. In contrast, the sequential auction gives final information about prices and assignments for all prior auctions. This final information may be more useful to bidders than the preliminary information revealed in a simultaneous auction.

Supporters of sequential auctions also point out that the great flexibility of a simultaneous auction makes it more susceptible to collusive strategies. Since nothing is assigned until the end in a simultaneous auction, bidders can punish aggressive bidding by raising the bids on those licenses desired by the aggressive bidder. In a sequential auction, collusion is more difficult. A bidder that is supposed to win a later license at a low price is vulnerable to competition from another that won an earlier license at a low price. The early winner no longer has an incentive to hold back in the later auctions.

In the end, the decision-makers at the FCC were convinced that the virtues of the simultaneous auction – greater information release and greater bidder flexibility in responding to information – would improve efficiency. Although the

FCC was concerned that a simultaneous auction might be more collusive, it felt that the setting was otherwise not conducive to collusion. In any event, sequential auctions would not eliminate the possibility for collusion.

The ability to successfully implement a novel auction form was a chief concern. However, the FCC was able to test and refine the simultaneous auction in the simpler and lower stake setting of narrowband licenses. In addition, experimental tests of the design were conducted before the first narrowband auction began. Charles Plott, David Porter, and John Ledyard conducted tests at CalTech's experimental lab. Several large bidders conducted their own tests as well. These tests provided evidence that the design would work in practice.

3.3. Package bids are too complex

A bidder's value of a license may depend on what other licenses it wins. Philadelphia may be worth more to a bidder if it wins the adjacent licenses in New York and Washington. Hence, bidders may value being able to bid on a combination of licenses, rather than having to place a number of individual bids. With a package bid, the bidder either gets the entire combination or nothing. There is no possibility that the bidder will end up winning just some of what it needs.

With individual bids, bidding for a synergistic combination is risky. The bidder may fail to acquire key pieces of the desired combination, but pay prices based on the synergistic gain. Alternatively, the bidder may be forced to bid beyond its valuation in order to secure the synergies and reduce its loss from being stuck with the dogs. This is the exposure problem. Individual bidding exposes bidders seeking synergistic combinations to aggregation risk.

Not allowing package bids can create inefficiencies. For example, suppose that there are two bidders for two adjacent parking spaces. One bidder with a car and a trailer requires both spaces. She values the two spots together at \$100 and a single spot is worth nothing; the spots are perfect complements. The second bidder has a car, but no trailer. Either spot is worth \$75, as is the combination; the spots are perfect substitutes. Note that the efficient outcome is for the first bidder to get both spots for a social gain of \$100, rather than \$75 if the second bidder gets a spot. Yet any attempt by the first bidder to win the spaces is foolhardy. The first bidder would have to pay at least \$150 for the spaces, since the second bidder will bid up to \$75 for either one. Alternatively, if the first bidder drops out early, she will 'win' one license, losing an amount equal to her highest bid. The only equilibrium is for the second bidder to win a single spot by placing the minimum bid. The outcome is inefficient, and fails to generate revenue. In contrast if package bids are allowed, then the outcome is efficient. The first bidder wins both licenses with a bid of \$75 for both spots.

This example is extreme to illustrate the exposure problem. The inefficiency involves large bidder-specific complementarities and a lack of competition. In

the PCS auctions and most other spectrum auctions, the complementarities are less extreme and the competition is greater.

Unfortunately, allowing package bids creates other problems. Package bids may favour bidders seeking large aggregations due to a variant of the free-rider problem, called the threshold problem (Bykowsky et al., 2000; Milgrom, 2000). Continuing with the last example, suppose that there is a third bidder who values either spot at \$40. Then the efficient outcome is for the individual bidders to win both spots for a social gain of $75 + 40 = \$115$. But this outcome may not occur when values are privately known. Suppose that the second and third bidders have placed individual bids of \$35 on the two licenses, but these bids are topped by a package bid of \$90 from the first bidder. Each bidder hopes that the other will bid higher to top the package bid. The second bidder has an incentive to understate his willingness to push the bidding higher. He may refrain from bidding, counting on the third bidder to break the threshold of \$90. Since the third bidder cannot come through, the auction ends with the first bidder winning both spaces for \$90.

A second problem with allowing package bids is complexity. If all combinations are allowed, even identifying the revenue maximising assignment is an intractable integer programming problem when there are many bidders and licenses. The problem can be made tractable by restricting the set of allowable combinations (Rothkopf et al., 1998). However, these restrictions may eliminate many desirable combinations, especially in broadband PCS where cellular holdings and other existing infrastructure tend to create idiosyncratic license synergies. Alternatively, a bid mechanism can be used that puts the computational burden on the bidders. In the AUSM system, bidders must propose bids that in combination with other bids exceed the amount bid for standing package bids (Banks et al., 1989).

Increased complexity is a legitimate concern when considering package bids. Although simultaneous auctions with package bids were successfully used in the laboratory (Bykowsky et al., 2000), it was far from certain that the FCC could successfully run auctions with package bids under the tight time schedule in the early auctions.

The FCC decided against allowing package bids in its early auctions. The threshold problem and increased complexity of package bids were thought to be worse than the exposure problem. However, since the initial PCS auctions, the FCC has done further research on package bidding, and intends to use it in the 700 MHz auction as discussed below.

3.4. Other issues

Having settled on a simultaneous ascending auction without package bids, several issues of implementation remained.

How much information should the bidders be given? The insights from Milgrom and Weber (1982) suggest that typically more public information is better. Hence, with the exception of bidder identities in the nationwide auction,

the FCC decided to reveal all information: the identities of the bidders, all the bids, and the bidders' current eligibility. So long as collusion and predatory bidding are not problems, revealing more information should improve efficiency and increase revenues. It also makes for a more open process.

Should the rounds be discrete or continuous? The FCC decided on discrete rounds, which would give the bidders a specific amount of time to respond to bids. Continuous bidding has the advantage that it makes endogenous the time between bids. Bidders can respond quickly when the strategic situation is simple, and take more time when it is complex. Discrete bidding is easier to implement and it gives the bidders a specific schedule to follow. Bidders know exactly when new information will be available and when they have to respond.

How can the FCC best control the pace of the auction? There are three key instruments: the frequency of rounds, the minimum bid increments, and an activity rule, which sets minimum levels of bidding activity. These are discussed later.

4. Simultaneous ascending auction

The simultaneous ascending auction has emerged as the standard approach to spectrum auctions. This design has been successfully used and refined by the FCC in over two-dozen auctions. Other countries, such as Australia, Canada, Mexico, the Netherlands, and the United Kingdom, have used this design as well. Below is a description of the rules for a generic FCC auction.

The simultaneous ascending auction works as follows. A group of licenses with strong value interdependencies are up for auction at one time. A bidder can bid on any collection of licenses in any round, subject to an activity rule which determines the bidder's current eligibility. The auction ends when a round passes with no new bids on any license. This auction form was thought to give the bidders flexibility in expressing values and building license aggregations. An auction winner gains the exclusive right to use the spectrum in accordance with FCC rules for a period of ten years. Licenses typically are renewed with a negligible charge provided the licensee has adhered to FCC rules and met buildout requirements. Licensees at any time may resell licenses purchased without special preference. Resale of licenses purchased with special preference is restricted to prevent 'unjust enrichment.'

Spectrum cap. To promote competition, a firm is limited in the quantity of spectrum it can hold in any market. For example in U.S. auctions, firms can hold no more than 45 MHz of broadband spectrum in any area, assuring that there are at least five broadband wireless competitors in each market.

Payment rules. Payments are received by the FCC at three times: (1) an upfront payment before the bidding begins assures that the bidder is serious. Any withdrawal or default penalties are taken from the bidder's upfront payment;

(2) down payment of 20 percent is paid within five business days after the close of the auction; (3) a final payment of the remaining 80 percent is paid within five business days of the award of the license. Licenses are awarded one to three months after the auction.

The upfront payment is a refundable deposit. It is due two weeks before the auction begins and defines the bidder's maximum eligibility in any round of bidding. A bidder interested in winning a large quantity of licenses would have to submit a large upfront payment; a bidder with more limited interests could submit a smaller upfront payment. The size of a license typically is measured in MHz-pop: the bandwidth in megahertz times the population in the license area. In early PCS auctions, each bidder made an upfront payment of \$.02 per MHz-pop for the largest combination of licenses the bidder intended to be active on in any round. A bidder is active on a license if it places a new bid in the round or was the high bidder on the license in the prior round. Thus, an upfront payment of \$6 million in the AB broadband PCS auction would make a bidder eligible to be active on 30 MHz licenses covering $6/(30\times.02) = 10$ million people. The upfront payment is not license specific; it simply limits total bidding activity in any round.

Designated entities. To encourage broad participation in wireless communications, designated firms (women, minorities, and/or small businesses) were given bidding credits on specific licenses. These credits, ranging from 10 percent to 40 percent, were intended to offset any disadvantage these firms faced in raising capital and providing services. Some auctions also gave designated entities favourable installment payments terms. These were later abandoned.

Minimum bid increments. To assure that the auction concludes in a reasonable amount of time, the FCC specifies minimum bid increments between rounds. Bid increments are set at the greater of a percentage increment and an absolute increment. Bid increments are adjusted in response to bidder behaviour. In the early rounds, when bid activity is high, the FCC sets larger bid increments; in the later rounds, when bid activity is low, the FCC sets smaller bid increments. Typically, the bid increments are between 5 and 20 percent.

Activity rule. The activity rule, proposed by Paul Milgrom and Robert Wilson, is a further device for controlling the pace of the auction. It forces a bidder to maintain a minimum level of activity to preserve its current eligibility. As the auction progresses, the activity requirement increases in stages. For example, the activity requirement might be 60 percent in stage 1, 80 percent in stage 2, and 100 percent in stage 3 (the final stage). With a 60 percent activity requirement, each bidder must be active on a quantity of licenses, measured in MHz-pop, equal to at least 60 percent of the bidder's current eligibility. If activity falls below the 60 percent level, then the bidder's current eligibility is reduced to its current activity divided by 60 percent. With a 100 percent activity requirement, the bidder must be active on 100 percent of its current eligibility or its eligibility drops to its current activity. The lower activity requirement early in the auction gives the bidder greater flexibility in shifting among license

aggregations early on when there is the most uncertainty about what will be obtainable.

A waiver prevents a reduction in eligibility in the event of bidder error or some other problem. Bidders typically are given five waivers. Waivers are applied automatically. An automatic waiver is used whenever a bidder's eligibility would otherwise fall as a result of its reduced bid activity. A bidder that does not wish to maintain its eligibility from the prior round may override the automatic waiver.

Number of rounds per day. A final means of controlling the pace of the auction is the number of rounds per day. Typically, fewer rounds per day are conducted early in the auction when the most learning occurs. In the later rounds, there is much less bidding activity, and the rounds can occur more quickly. The FCC has conducted as few as one round per day and as many as twenty per day.

Stopping rule. A simultaneous stopping rule is used to give the bidders maximum flexibility in pursuing backup strategies. All markets close if a single round passes in which no new bids are submitted on any license, and the auction is in its final stage.

Bid information. Each bidder is fully informed about the identities of the bidders, the size of the upfront payments, and which bidders qualify as designated entities. High bids and bidder identities are posted after each round. In addition, all bids and bidder identities are displayed at the conclusion of each round, together with each bidder's eligibility and waivers.

Bid withdrawal. The high bidders can withdraw their bids subject to a bid withdrawal penalty. If a bidder withdraws its high bid, the FCC is listed as the high bidder and the minimum bid is the second-highest bid for that license. The second-highest bidder is in no way responsible for the bid, since this bidder may have moved on to other properties. If no firm bids on the license, the FCC can reduce the minimum bid. Typically, the FCC drops the minimum bid at most once, before committing not to reduce the minimum bid further.

To discourage insincere bidding, there are penalties for withdrawing a high bid. The penalty is the larger of 0 and the difference between the withdrawn bid and the final sale price. This penalty is consistent with the standard remedy for breach of contract. The penalty equals the damage suffered by the FCC as a result of the withdrawal. If the bidder defaults or is disqualified after the close of the auction, the penalty is increased by 3 percent of the eventual sale price to compensate the FCC for additional selling costs. The additional 3 percent default payment is also intended to discourage defaults (after the auction closes) relative to withdrawals (during an auction).

Since we do not observe the values firms place on licenses, it is impossible to directly assess the efficiency of the simultaneous ascending auction. Nonetheless, we can indirectly evaluate the auction design from the observed behaviour (Cramton, 1998b; McAfee and McMillan, 1996). I focus especially on the three PCS broadband auctions, which I refer to as the AB auction, the C auction, and the DEF auction, indicating the block(s) of spectrum that were assigned in each.

Revenue is a first sign of success. Auction revenues have been substantial. Revenues in US auctions typically have exceeded industry and government estimates. The simultaneous ascending auction may be partially responsible for the large revenues. By revealing information in the auction process, the winner's curse is reduced, and the bidders can bid more aggressively. Also, revenues may increase to the extent the design enables bidders to piece together more efficient packages of licenses.

A second indicator of success is that the auctions tended to generate market prices. Similar items sold for similar prices. In the narrowband auctions, the price differences among similar licenses were at most a few percent and often zero. In the first broadband auction, where two licenses were sold in each market, the prices differed by less than one minimum bid increment in 42 of the 48 markets.

A third indicator of success is the formation of efficient license aggregations. Bidders did appear to piece together sensible license aggregations. This is clearest in the narrowband auctions. In the nationwide narrowband auction, bidders buying multiple bands preferred adjacent bands. The adjacency means that the buffer between bands can be used for transmission, thus increasing capacity. The two bidders that won multiple licenses were successful in buying adjacent bands. In the regional narrowband auction, the aggregation problem was more complicated. Several bidders had nationwide interests, and these bidders would have to piece together a license in each of the five regions, preferably all on the same band, in order to achieve a nationwide market. The bidders were remarkably successful in achieving these aggregations. Four of the six bands sold as nationwide aggregations. Bidders were able to win all five regions within the same band. Even in the two bands that were not sold as nationwide aggregations, bidders winning multiple licenses won geographically adjacent licenses within the same band.

Large aggregations were also formed in the PCS broadband auctions. Bidders tended to win the same band when acquiring adjacent licenses. In the AB auction, the three bidders with nationwide interests appeared to have efficient geographic coverage when one includes their cellular holdings. The footprints of smaller bidders also seem consistent with the bidders' existing infrastructures. In the C-block auction, bidders were able to piece together contiguous footprints, although many bidders were interested in stand-alone markets.

Ausubel et al. (1997) and Moreton and Spiller (1998) analyse the AB and C auction data to see if there is evidence of local synergies. Consistent with local synergies, these studies find that bidders did pay more when competing with a bidder holding neighbouring licenses. Hence, bidders did bid for synergistic gains and, judging by the final footprints, often obtained them.

The two essential features of the FCC auction design are (1) the use of multiple rounds, rather than a single sealed bid, and (2) simultaneous, rather than sequential sales. The goal of both of these features is to reveal information and then give the bidders the flexibility to respond to the information. There is

substantial evidence that the auction was successful in revealing extensive information. Bidders had good information about both prices and assignments at a point in the auction where they had the flexibility to act on the information (Cramton, 1997). The probability that a high bidder would eventually win the market was high at the midpoint of each auction. Also the correlation between mid-auction and final prices was high in each auction. Information about prices and assignments improved throughout each auction and was of high quality before bidders lost the flexibility to move to alternative packages.

The absence of resale also suggests that the auctions were highly efficient. In the first two years of the PCS auctions, there was little resale. GTE was the one exception. Shortly after the AB auction ended, GTE sold its AB winnings for about what it paid for the licenses. Apparently there was a shift in corporate strategy away from PCS and toward cellular.

5. Demand reduction and collusive bidding

Despite the apparent success of these auctions, an important issue limiting both efficiency and revenues is demand reduction and collusive bidding. This issue stems from the fact that these are multiple-item auctions. The efficiency results from single-item auctions do not carry forward to the multiple-item setting. In an ascending auction for a single item, each bidder has a dominant strategy of bidding up to its private valuation. Hence, the item always goes to the bidder with the highest value. If, instead, two identical items are being sold in a simultaneous ascending auction, then a bidder has an incentive to stop bidding for the second item before its marginal valuation is reached. Continuing to bid for two items raises the price paid for the first. As a result, the bidder with the highest value for the second item may be outbid by a bidder demanding just a single unit.

This logic is quite general. In multi-unit uniform-price auctions, typically every equilibrium is inefficient (Ausubel and Cramton, 1996). Bidders have an incentive to shade their bids for multiple units, and the incentive to shade increases with the quantity being demanded. Hence, large bidders will shade more than small bidders. This differential shading creates an inefficiency. The small bidders will tend to inefficiently win licenses that should be won by the large bidders. The intuition for this result is analogous to why a monopolist's marginal revenue curve lies below its demand curve: bringing more items to market reduces the price received on all items. In the auction, demanding more items raises the price paid on all items. Hence, the incentive to reduce demand.

The FCC spectrum auctions can be viewed as an ascending-bid version of a uniform-price auction. Certainly, for licenses that are close substitutes, the simultaneous ascending auction has generated near uniform prices for similar items. However, the incentives for demand reduction and collusive bidding likely are more pronounced in an ascending version of the uniform-price auction

(Cramton and Schwartz, 1999; Ausubel and Schwartz, 1999)[1]. To illustrate this, consider a simple example with two identical goods and two risk-neutral bidders. Suppose that to each bidder the marginal value of winning one item is the same as the marginal value of winning a second item. These values are assumed independent and private, with each bidder drawing its marginal value from a uniform distribution on [0, 100]. First consider the sealed-bid uniform price auction where each bidder privately submits two bids and the highest two bids secure units at a per-unit charge equal to the third highest bid. There are two equilibria to this sealed-bid auction: a demand-reducing equilibrium where each bidder submits one bid for \$0 and one bid equal to its marginal value; and a sincere equilibrium where each bidder submits two bids equal to its marginal value. The sincere equilibrium is fully efficient in that both units will be awarded to the bidder who values them more. The demand-reducing equilibrium, however, raises zero revenue (the third-highest bid is zero) and is inefficient since the bidder with the higher value wins only one unit.

Next consider the same setting, but where an ascending version of the auction is used. Specifically, view the ascending auction as a two-button auction where there is a price clock that starting from price 0 increases continuously to 100. The bidders depress the buttons to indicate the quantity they are bidding for at every instant. The buttons are 'non-repushable' meaning a bidder can decrease its demand but cannot increase its demand. Each bidder observes the price and can observe how many buttons are being depressed by its opponent. The auction ends at the first price such that the total number of buttons depressed is less than or equal to two. This price is called the stop-out price. Each bidder will win the number of units she demands when the auction ends, and is charged the stop-out price for each unit she wins. In this game, if weakly dominated strategies are eliminated, there is a unique equilibrium in which the bidding ends at a price of zero, with both bidders demanding just a single unit. The reason is that each bidder knows that if it unilaterally decreases its bidding to one unit, the other bidder will instantaneously end the auction, as argued above. But since the bidder prefers the payoff from winning one unit at the low price over its expected payoff of winning two units at the price high enough to eliminate the other bidder from the auction, the bidder will immediately bid for just one unit, inducing an *immediate* end to the auction. Thus, the only equilibrium here is analogous to the demand-reducing equilibrium in the sealed-bid uniform-price auction. The efficient equilibrium does not obtain. This example shows that the incentives to reduce demand can be more pronounced in an open auction, where bidders have the opportunity to respond to the elapsed bidding. The 1999 German GSM spectrum auction, which lasted just two rounds, illustrates this behaviour (Jehiel and Moldovanu, 2000).

[1] But see Perry and Reny (1999) that construct an efficient equilibrium in the simultaneous ascending auction.

This example is meant to illustrate that in simple settings with few goods and few bidders, bidders have the incentive to reduce demand. Direct evidence of demand reduction was seen in the nationwide narrowband auction. The largest bidder, PageNet, reduced its demand from three of the large licenses to two, at a point when prices were still well below its marginal valuation for the third unit (Cramton, 1995). PageNet felt that, if it continued to demand a third license, it would drive up the prices on all the others to disadvantageously high levels.

An examination of the bidding in the AB auction is suggestive that the largest bidders did drop out of certain markets at prices well below plausible values, as a result of either demand reduction or tacit collusion.

Further evidence of demand reduction comes from the C auction. One large bidder defaulted on the down payment, so the FCC re-auctioned the licenses. Interestingly, the licenses sold for 3 percent more than in the original auction. Consistent with demand reduction, NextWave, the largest winner in the C auction, bought 60 percent of the re-auctioned spectrum. This occurred despite the fact that NextWave was not the second-highest bidder on any of these licenses in the original auction. NextWave was able to bid aggressively in the re-auction, knowing that its bidding would have no affect on prices in the original auction.

Engelbrecht-Wiggans and Kahn (1999) and Brusco and Lopomo (1999) show that for an auction format like the FCC's, where the bidding occurs in rounds and bidding can be done on distinct units, that there exist equilibria where bidders co-ordinate a division of the available units at low prices relative to own values. Bidders achieve these low-revenue equilibria by threatening to punish those bidders who deviate from the cooperative division of the units. The idea in these papers is that bidders have the incentives to split up the available units ending the auction at low prices. With heterogeneous goods and asymmetric bidders in terms of budgets, capacities, and current holdings of complementary goods, it is unlikely that bidders would be aware of a simple equilibrium strategy that indicates which licenses to bid on and which to avoid. However, bidders in the FCC auctions, especially the DEF auction, took advantage of signalling opportunities to co-ordinate how to assign the licenses. With signalling, bidders could indicate which licenses they most wanted and which licenses they would be willing to forgo. Often this communication took the form of punishments.

Cramton and Schwartz (1999) examine collusive bidding strategies in the DEF auction. During the DEF auction the FCC and the Department of Justice observed that some bidders used bid signalling to co-ordinate the assignment of licenses. Specifically, some bidders engaged in *code bidding*. A code bid uses the trailing digits of the bid to tell other bidders on which licenses to bid or not bid. Since bids were often in millions of dollars, yet were specified in dollars, bidders at negligible cost could use the last three digits – the trailing digits – to specify a market number. Often, a bidder (the sender) would use these code bids as retaliation against another bidder (the receiver) who was bidding on a license desired

by the sender. The sender would raise the price on some market the receiver wanted, and use the trailing digits to tell the receiver on which license to cease bidding. Although the trailing digits are useful in making clear which market the receiver is to avoid, *retaliating bids* without the trailing digits can also send a clear message. The concern of the FCC is that this type of coordination may be collusive and may dampen revenues or efficiency.

The DEF auction was especially vulnerable to collusive bidding, it featured both small markets and light competition. Small markets enhanced the scope for splitting up the licenses. Light competition increased the possibility that collusive bidding strategies would be successful. Indeed, prices in the DEF auction were much lower than prices in the two earlier broadband PCS auctions.

From a strategic viewpoint, the simultaneous ascending auction can be thought of as a negotiation among the bidders. The bidders are negotiating how to split up the licenses among themselves, but only can use their bids for communication. The auction ends when the bidders agree on the division of the licenses. Retaliating bids and code bids are strategies to co-ordinate on a split of the licenses at low prices. In addition, bidders with a reputation for retaliation may scare off potential competitors. Cramton and Schwartz (CS) hypothesise that bidders who commonly use these strategies pay less for the spectrum they ultimately win.

CS find that six of the 153 bidders in the DEF auction regularly signalled using code bids or retaliating bids. These bidders won 476 of the 1,479 licenses for sale in the auction, or about 40 percent of the available spectrum in terms of population covered. Controlling for market characteristics, these signalling bidders paid significantly less for their licenses.

Further evidence that retaliation was effective in reducing prices is seen by the absence of arbitrage between the D and E blocks in each market. In particular, CS find that there was a tendency for bidders to avoid AT&T, a large bidder with a reputation for retaliation. If bidders did not care about the identity of the high bidder, they would arbitrage the prices of the D and E blocks, and bid against AT&T if the other block was more expensive. This did not happen. Even when the price of the other block was 50 percent higher, bidders bid on the higher priced block 27 percent of the time, rather than bid against AT&T.

Following the experience in the DEF auction, the FCC restricted bids to a whole number of bid increments (typically between 1 and 9) above the standing high bid. This eliminates code bidding, but it does nothing to prevent retaliating bids. Retaliating bids may be just as effective as code bids in signalling a split of the licenses, when competition is weak.

The auctioneer has many instruments to reduce the effectiveness of bid signalling. These include:

- Concealing bidder identities. This prevents the use of targeted punishments against rivals. Unless there are strong efficiency reasons for revealing identities, anonymous auctions may be preferable.

- Setting high reserve prices. High reserve prices reduce the incentive for demand reduction in a multiple-item auction, since as the reserve price increases the benefit from reducing demands falls. Moreover, higher reserve prices reduce the number of rounds that the bidders have to co-ordinate a split of the licenses and still face low prices.
- Offering preferences for small businesses and non-incumbents. Competition is encouraged by favouring bidders that may otherwise be disadvantaged *ex ante*. In the DEF auction, competition for the D and E license could have been increased by extending small business preferences to the D and E blocks, rather than restricting the preferences to the F block.
- Offering larger licenses. Many small licenses are more easily split up. At the other extreme a single nationwide license is impossible to split up. In the absence of synergies, such extreme bundling may have negative efficiency consequences, but improve revenues.

In auctions for identical items, the inefficiencies of demand reduction can be eliminated with a Vickrey (1961) auction. Alternatively, one can use Ausubel's (1997) ascending implementation of the static Vickrey auction, which has the additional advantages of an ascending-bid design. However, most spectrum auctions are not for identical items, so Vickrey-type mechanisms often are not practical.

6. Lessons learned and auction enhancements

The FCC auction rules have evolved in response to the experience of more than two dozen auctions. An examination of this evolution is instructive. Despite many enhancements, the FCC spectrum auctions have retained the same basic structure, a strong indication of an excellent initial design. The intent of the changes was to reduce speculative bidding, to avoid collusion, and to speed the auction along.

Elimination of installment payments. A potentially serious inefficiency in the C auction was speculative bidding caused by overly attractive installment payment terms. Bidders only had to put down 5 percent of their bids at the end of the auction, a second 5 percent at the time of license award, and then quarterly installment payments at the 10-year Treasury rate with interest-only payments for the first 6 years. These attractive terms favour bidders that are speculating in spectrum. If prices go up, the speculators do well; if prices fall, the speculators can walk away from their down payments. Indeed, spectrum prices did fall after the C auction, and most of the large bidders in the C auction defaulted on the payments. As a result of this experience, the FCC no longer offers installment payments. Bids must be paid in full when the licenses are awarded.

Click-box bidding. Bidders in FCC auctions no longer enter bids in dollars. Rather, the bidder indicates in a click-box the number of bid increments from 1–9 that it wishes to bid above the standing high bid. If the standing high bid is 100 and the minimum bid increment is 10 percent, then the allowable bids would be

110, 120, . . . ,190, corresponding to the allowable increment bids of 1, 2, . . . , 9. This approach solves two problems. First, it eliminates code bidding. Bidders can no longer use the trailing digits of bids to signal to other bidders who should win what. Second, it reduces the possibility of mistaken bids. There were several instances of bidders adding too many zeros to the end of their dollar bids. With click-box bidding, substantial jump bids are permitted but not gross mistakes.

The downside of click-box bidding is the greater possibility of tie bids. This turns out not to be a serious problem. Although ties do occur early in the auction, it is unlikely that the final bid on a license involves a tie. Still ties do occur, and so the FCC tie-breaking rule takes on greater importance. The FCC breaks ties with the time stamp. Bids entered earlier have preference. Since bidders often have a mild preference for being the standing high bidder, it is common for bidders to race to enter their bids early in the round to win ties. Such behaviour is undesirable, and so the FCC is now considering a random tie-breaking rule in future auctions.

License-specific bid increments. In early auctions, the FCC used the same percentage increment for all licenses. This was fine for auctioning a handful of similar licenses. However, when auctioning hundreds of heterogeneous licenses, it was found that some licenses would have a lot of bidding activity and others would have little activity. To speed the auction along, it makes sense to use larger bid increments for more active licenses. In recent auctions, the FCC adjusts the bid increments for each license based on the licence's history of bid activity, using an exponential smoothing formula. Percentage increments tend to range between 5 and 20 percent, depending on prior activity. More active licenses have a larger increment.

Limit the use of withdrawals. Bid withdrawals were introduced to permit bidders to back out of a failed aggregation. The DEF auction had 789 withdrawals. Few if any of these withdrawals were related to failed aggregations. Rather, most of the withdrawals appear to have been used as a strategic device, in one of two ways: (1) as a signal of the bidder's willingness to give up one license in exchange for another or (2) as part of a parking strategy to maintain eligibility without bidding seriously. This undesirable use of withdrawals was also observed in other auctions. As a result, the FCC now only allows withdrawals in at most two rounds of the auction for any bidder. This enables the bidder to back out of up to two failed aggregations, and yet prevents the frequent strategic use of withdrawals.

Combine bid submission and withdrawal phases. The FCC originally divided a round of bidding into two phases, a bidding phase followed by a withdrawal phase. The separation of these phases was largely a historical artefact and provides little benefit. Its main effect was to impede the pace of the auction.

The FCC has since combined the two phases into one. With the combined procedure, bid submission consists of two steps: withdrawal followed by submission. In the withdrawal step, the bidder may withdraw on any or all licenses on which it is the high bidder. Then, in the bid submission step, the bidder places

any desired new bids, with available eligibility increased to reflect the withdrawals. Hence, a bidder withdrawing in New York can then place a bid (in the same round) on Los Angeles, because of the eligibility freed by the New York withdrawal.

Faster rounds. The FCC's auction system now permits much faster rounds than the initial implementation. In many auctions, bidding activity is slow in the later part of the auction. Hence, being able to conduct 20 or more rounds per day is important in speeding the auction along.

Minimum opening bids. Early FCC auctions did not use minimum opening bids; any opening bid greater than zero was acceptable. The FCC now sets substantial minimum opening bids. These bid limits both increase the pace and reduce the potential for collusion. By starting at a reasonably high level, the bidders have fewer rounds to resolve their conflicts at low prices. The benefit of collusive strategies is reduced.

7. Package bidding

One auction enhancement that has received considerable attention is package bidding. A key simplification of the FCC auctions is only allowing bids on individual licenses. This works well in settings where there are not large complementarities across licenses, or if there are large complementarities, then where the complementarities are similar among the bidders. In such a setting, the exposure problem is slight. There is little likelihood that a sincere bidder will get stuck with licenses it does not want. However, in auctions where license complementarities are large and differ among bidders, then the exposure problem may be substantial. In this latter case, allowing package bids – all-or-nothing bids on collections of complementary licenses – can reduce the exposure problem and improve efficiency. Efficiency can also be improved by reducing the incentives for demand reduction by large bidders.

Auctions with package bidding, often referred to as combinatorial auctions or package auctions, are actively being researched. In May 2000, the FCC sponsored a conference on the topic[2]. A goal of the conference was to determine the best way to introduce package bidding in the 700 MHz auction to take place in Fall 2000. The 700 MHz auction represented a good test-case for package bidding for two reasons. First, it is a relatively simple case, since it involves only 12 licenses: 2 bands (one 10 MHz and one 20 MHz) in each of 6 regions. Second, prospective bidders had expressed interest in alternative packaging. Those bidders intending to provide a fixed high-speed data service desired the full 30 MHz in a particular region. Some mobile wireless providers desired individual licenses to add capacity

[2] The conference materials are available at http://combin.fcc.gov/. This site also includes the comments, reply comments, and FCC documents on package bidding.

in congested regions or to fill out their footprint. Still others desired nationwide packages to offer a nationwide mobile service.

The conference resulted in FCC Public Notice DA00-1075 seeking comment on modifying the simultaneous ascending auction to allow package bidding in the 700 MHz auction. After comments and reply comments were received, the FCC issued Public Notice DA00-1486 adopting and describing the package bidding rules for the 700 MHz auction. I briefly describe the approach taken by the FCC in this auction. The rules are intended to improve the efficiency of the auction by avoiding the exposure problem, while limiting the threshold problem.

A bidder can bid on the individual licenses as in the standard simultaneous ascending auction. In addition, a bidder can place all-or-nothing bids on up to twelve packages, which the bidder determines at any point in the auction. In this way the bidder can avoid the exposure problem when licenses are complements. The provisional winning bids are the set of consistent bids that maximise total revenues. Consistent bids are bids that (1) do not overlap, and (2) are made or renewed in the same round. Bids made by a bidder in different rounds are treated as mutually exclusive[3].

Limiting each bidder to twelve bidder-specific packages simplifies the auction, and still gives the bidders great flexibility in expressing synergies for various license combinations. The FCC's original proposal allowed only nine package bids: the six 30 MHz regional bids, and three nationwide bids (10, 20, or 30 MHz). Although these nine packages were consistent with the expressed desires of many perspective bidders, others felt that the nine packages were too restrictive.

The activity rule is unchanged, aside from a new definition of activity and a lower activity requirement of 50 percent, giving the bidders greater flexibility in shifting among packages. A bidder must be active on 50 percent of its current eligibility or its eligibility in the next round will be reduced to two times its activity. A bid counts as activity if (1) it is part of a provisionally winning set in the prior round, or (2) it is a new bid or a renewal of a provisionally winning bid in the prior round. A bidder's activity level is the maximum number of bidding units a bidder can win considering only those licenses and packages on which the bidder is active. Bids made in different rounds are treated as mutually exclusive; hence, a bidder wishing to add a license or package to its provisional winnings must renew the provisional winning bids in the current round.

The FCC adopted a two-round simultaneous stopping rule. The auction ends after two consecutive rounds of no new bids on any licenses or packages. Renewed bids do not count as new bids. The two-round approach gives bidders fair warning that the auction may end if they do not place a new bid.

The determination of minimum bids is an important change in the rules, since it impacts the extent of the threshold problem faced by those bidding on individual

[3] This and several other features of the design were recommended by Paul Milgrom. See http://wireless.fcc.gov/auctions/31/.

licenses or small packages. The minimum bid on a license or package is the greater of: (1) the minimum opening bid, (2) the bidder's own previous high bid on that package plus x percent, and (3) the number of bidding units of the package times the lowest $/bidding unit on any package in the last five rounds. The FCC specifies x, and retains discretion to adjust minimum bids on a license-by-license or package-by-package basis.

The key feature of this minimum bid rule is point (2), which makes the minimum bid depend on the bidder's prior high bid for the package. This recognises that since a bidder's bids in different rounds are mutually exclusive, another bidder's bid on a package can be a provisionally winning bid even if it is less than the high bid of the prior round. Hence, the rules allow bids that are below another's prior high bid. Points (1) and (3) limit how low a bidder can bid when it starts out bidding on a new package. The bid must be at least as great as the minimum opening bid and the least expensive package on a per unit basis.

The FCC will continue to use 'click box' bidding, in which the bidder specifies either the minimum bids or an integer between 1 and 9. A bid of 1 is the minimum bid plus x percent of the minimum bid; a bid of 2 is the minimum bid plus 2x percent of the minimum bid; and so on. The FCC considered limiting bids on packages to the minimum bid to reduce the threshold problem; however, it was felt that the revised minimum bid rule adequately addressed the threshold problem.

Another change is allowing bidders to submit 'last and best' bids. Last and best bids can be for any amount between the bidder's prior high bid and the minimum bid. A bidder cannot bid again after placing one or more last and best bids.

A bidder can renew the highest bid it made on any license or package. Renewing a bid does not increase the bid amount. Renewed bids are needed, since bids in different rounds are treated as mutually exclusive. Thus, if a bidder wishes to win both its provisional winners from the prior round plus a new license, it must bid on the new license and renew the provisional winners. Activity credit is not given for renewing a bid that is not a provisional winner.

Provisional winning bids are the set of 'consistent' bids (bids that do not overlap and are made or renewed by a bidder in the same round) that maximise revenues as of the current round. Winning bids are the provisional winning bids at the end of the auction. Ties are broken randomly. Licenses on which no bids have been submitted are treated as if the minimum opening bid was placed. This is consistent with the minimum opening bids reflecting the FCC's opportunity cost of selling the licenses.

Treating a bidder's bids in different rounds as mutually exclusive is an important feature. First, it gives bidders a vehicle for submitting contingent 'or' bids. If the bidder desires one package or the other, but not both, the bidder simply bids for the two packages in separate rounds. Bidders can shift to backup strategies without fear that they will win a collection of licenses they do not desire. Second, it encourages sincere bidding early in the auction, since bids placed early in the auction may emerge as winning bids. When submitting its bids in a round,

the bidder is saying that it is happy to win any of the submitted bids, including renewed bids, at the prices bid.

Bid withdrawals are not permitted. In an auction without package bidding, withdrawals were needed as a device for backing out of a failed aggregation. With package bidding, withdrawals are not needed. There is no exposure problem, since the bidder can bid all-or-nothing on a package of complementary licenses.

Allowing package bids is a major change from the original FCC design. Bidder incentives are fundamentally altered. Large bidders, in addition to not facing an exposure problem, have less of an incentive for demand reduction. Small bidders now face a negotiation with each other on how to top large package bids. All bidders will need to rethink their strategies. Relative to the standard auction, which favoured small bidders, the package auction favours large bidders.

The FCC's package auction is an important innovation. Implementing such an auction poses many challenges for the FCC. It is not surprising that it took several years to introduce package bidding. Even now it is uncertain how well this design will perform in practice compared with the standard auction without package bidding. Whether the package auction is an improvement surely depends on the setting.

8. UMTS auctions in Europe and Asia

European and Asian countries currently are in the process of assigning Universal Mobile Telecommunications System (UMTS) licenses, which will be used for third-generation mobile wireless services. Many countries decided to use an auction to assign the licenses (e.g., U.K., Netherlands, Germany, and Switzerland); other countries (e.g., France, Spain, and Sweden) decided on a beauty contest; and still others (e.g., Italy) decided on a hybrid beauty contest/auction. The allocation of UMTS spectrum is European-wide, although individual countries can determine there own band plans. Nearly all countries are assigning between four and six national licenses.

The early European auctions conducted in 2000 raised nearly $100 billion. An additional expense of at least $100 billion will be required to build the infrastructure necessary to provide service. In countries like the U.K. and Germany, which fetched the highest prices the license cost per person amounts to about $100 per person. These staggering sums have led many to question the desirability of auctions for assigning spectrum licenses.

In fact, auction prices have varied considerably over time and over markets. This is seen in Figure 1, which presents the per person price of a 20 MHz license (2×10 MHz paired) in several major spectrum auctions[4]. For comparison

[4] Most of the European licenses also included 5 MHz unpaired spectrum. However, these auctions have shown that the bidders place little value on this unpaired spectrum, so I ignore the unpaired spectrum in the price comparison. In the US C-block auction, bidders received attractive installment payment plans. I discount these prices by 40 percent to reflect the value of the installment payments.

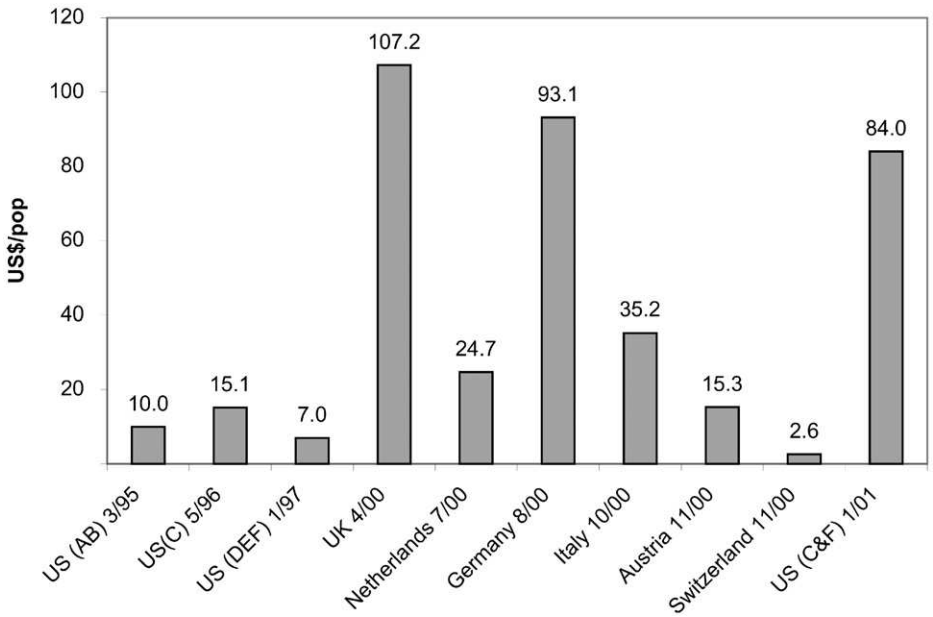


Fig. 1. Mobile wireless price comparison (2×10 MHz + 5 MHz).

purposes, Figure 1 also shows past and current US auctions of 2G spectrum. The first three US auctions occurred over three years before the 3G auctions in Europe. The fourth US auction concluded in January 2001 with a price comparable to the highest 3G prices. Part of price variation is explained by the different times at which the auctions occurred. Part of the difference in the European prices is explained by the size of the various countries. Markets like the U.K. and Germany are thought to have more value, even on a per person basis, than the Netherlands and Switzerland. Still there is much variation to explain. The primary determinant of prices appears to be the level of competition going into the auction, rather than the subtle differences in auction design across the various countries. Competition in the auction is largely endogenous, since it is the result of partnership negotiations among potential bidders.

The first countries to auction (U.K., Netherlands, and Germany) used a simultaneous ascending auction. In the U.K. and the Netherlands, the auction design is especially simple, since a bidder can win at most a single license. Hence, there is no incentive for demand reduction.

An incumbent provider of first- and second-generation mobile services in a market tends to have a much higher value for a license than a new entrant. The incumbent's existing infrastructure and customer base yields cost savings in

building out a service and attracting customers. Also, an incumbent's failure to win a license would adversely impact its existing business. Hence, it is likely that incumbents will win licenses. As a result, the outcome of the auction depends greatly on the number of licenses that are auctioned relative to the number of incumbents. The U.K. government initially intended to auction four licenses, but there was a concern that this would damage auction revenues, since potential new entrants might fear that it is pointless to compete against the four strong incumbents. The final design involved the auction of five licenses, which guaranteed that a new entrant would win a license. In contrast, in the Netherlands, there are five incumbents and five UMTS licenses.

The U.K. UMTS auction was the largest auction in history, generating $35 billion in revenues. Four incumbents and nine potential entrants competed in the auction, which lasted 150 rounds. The four incumbents, British Telecom, Vodafone, Orange, and One2One, each won a license, as did the new entrant TIW. Vodafone's 2×15 MHz license sold for $9.5 billion, or $160 per person; the other incumbent licenses sold for similar prices on a per MHz basis. This amount far exceeded industry expectations before the auction. In contrast, the average price paid in 1995 in the FCC's AB auction for a 2×15 MHz license was $15.5 per person, less than one-tenth of the U.K. price.

Why were the U.K. prices so high? There are several factors. First, the 1995 price of $15.5 per person in the US is out of date. US prices now are much higher. For example, early in 2000, Nextel offered to buy Nextwave's spectrum for $8.3 billion, a price of about $80 per person. Second, unlike in the US auction, there was no possibility for demand reduction. Bidders had every incentive to bid up to their values. Third and likely most important, the bidding was especially competitive, since this was the first in a sequence of UMTS auctions throughout Europe. A winner in the U.K. auction is well positioned for subsequent UMTS auctions. Hence, a bidder can view the U.K. auction as a foot-in-the-door to Europe. To the extent that a U.K. license is complementary with UMTS licenses in other countries, then we should expect the U.K. licenses to sell for a premium, recognising this complementary value.

In the German auction, each bidder bid for either two or three of the twelve available 2×5 MHz blocks; hence, there will be between four and six winners. The German auction had the apparent advantage that it lets the bidding determine how many UMTS competitors there will be. If competition among new entrants is intense, there can be six winners, presumably the four incumbents plus two new entrants. Jehiel and Moldovanu (2000) argue that the design is biased against new entry, since the incumbents have substantially higher values because of their incumbent position and benefit from excluding new entrants. However, it is possible that the two weakest incumbents may be willing to bid for just two blocks, making room for a new entrant. The question is whether the benefit from demand reduction more than compensates for the reduced profits in a five-player vs. a four-player market. Probably a better design would set-aside two blocks for

a new entrant and then let the bidding determine whether there would be four or five winners for the remaining ten blocks.

Following the $35 billion UMTS auction in the U.K., the Dutch government expected to raise $8.7 billion in its auction. The Dutch government arrived at this figure by simply scaling down the U.K. amount for the smaller population in the Netherlands (15 vs. 59 million people). Confident of a huge windfall, the government cancelled its July bond issue. Their confidence may be a mistake. The reality is that the U.K. and Dutch auctions are quite different despite the fact that both use the simultaneous ascending auction to auction off five third-generation mobile licenses.

In the U.K., the guaranteed success of at least one new entrant encouraged participation in the auction. In the Netherlands, with five incumbents bidding for five licenses, the logical outcome is for the five incumbents to win licenses. Recognising the difficulty of winning a license, potential entrants have strong incentives to partner with the incumbent bidders. This is exactly what happened. Although initially there were several strong potential entrants, all partnered with one of the incumbents before the auction began. The strongest entrant, Deutsche Telekom, partnered with the weakest incumbent, Ben; DoCoMo and Hutchinson partnered with KPN; and NTL was already effectively partnered with Dutchtone (France Telecom has a large interest in both). This left one weak entrant in the bidding. At the beginning of the auction, just six bidders were competing for five licenses: five strong incumbents and one weak potential entrant (Versatel).

To make matters worse, the government specified minimum opening bids that decline in the first three rounds if a license does not receive a bid. The two large licenses (15 MHz-paired) start at 100 million guilders, but fall to 75 after one round of inactivity, 35 after two rounds, and 0 after three rounds; similarly, the three 10 MHz-paired licenses start at 90, but drop to 60, 30, and then 0, with each round of inactivity. Bidders were given three waivers of the activity rule that otherwise would require them to be active in each round.

The strategic use of waivers in this environment should be obvious. Even if the other five bidders bid in the first round, it is extremely unlikely that each of the five licenses would receive a bid. Hence, by using a waiver in the first round, a bidder can get the minimum opening bid to drop substantially. In fact, five of the six bidders saw this strategy. Only Libertel bid in the opening round, placing a 100 million guilder bid on B. (Libertel is 70 percent owned by Vodafone, the company that won the B license in the U.K. auction.) In the second and third rounds, the five bidders that needed to bid again used waivers. So going into the fourth round, the minimum bids were 0 on all licenses, but the B license, which stood at 110 million guilders. After 175 rounds of bidding, the three 10 MHz licenses were priced at 1/8th of the minimum opening bid (11 million guilder).

The Dutch auction rules also did not include the customary language that the auctioneer has the right to cancel the auction at any time. If Versatel decides to

drop out at low prices the auction will end at only a tiny fraction of the $8.7 billion figure the government is hoping for.

The 3G auctions provide an excellent test of auction theory with multiple items. These auctions have been strategically simple and are extremely high stake. For example, the German and Austrian 3G auctions – as well as an earlier 2G German auction – provided relatively clean tests of demand reduction. The theoretical predictions on demand reduction were borne out empirically in the final outcomes of these auctions. At the same time, one would have imagined that all of the bidding behaviour, at least by the eventual winners, could be easily understood as equilibrium behaviour in a profit-maximising model. However, this does not always appear to be the case.

One explanation for non-profit maximising bidding is the principal-agent problem between the shareholders and the CEO making the bidding decisions (Ausubel and Cramton, 2001). At times the CEO may not fully represent shareholder interests. For example, CEO's may sometimes fail to fully appreciate the economics of demand reduction, and fail to reduce demand as early as they should. In particular, in the German auction, Deutsche Telekom (DT) was in a position to likely end the German auction by dropping from demanding three blocks to demanding two blocks, at well before the final prices. Instead, DT continued bidding for three blocks for a time, ultimately buying the two blocks that it could have bought earlier, but paying about $2 billion more. This bidding behaviour may have more to do with the preferences (and fears) of the CEO, rather than the interests of the shareholder.

A second source of distortion is that the CEO simply acts as humans do in auctions, making some of the same mistakes that are frequently observed in the laboratory (Kagel and Roth, 1995). An example similar to the DT example, but with a different outcome occurred in the US C&F auction. In New York, three licenses were up for bid. After round 14, only the three largest bidders (Verizon, Cingular, and AT&T) were competing for the three licenses. At this point prices were at $782 million. Verizon, however, continued to bid for two licenses throughout the auction, creating excess demand until Cingular finally dropped out when Verizon's price exceeded $2 billion per license. Why did Verizon continue to push the price up on both of its licenses rather than ending the bidding at a fraction of the eventual price? One explanation is that the Verizon CEO felt that he would look bad, pushing the prices up and then eventually settling for one license. This gives the CEO an incentive to 'throw good money after bad,' a phenomenon for which there is ample experimental support.

9. Advice to Governments

Important lessons about spectrum auction design can be gleaned from recent experience.

9.1. Allocating the spectrum is just as important as its assignment

There are two steps in making spectrum available to companies. The first step is the allocation of the spectrum for licensing. The allocation defines the license (the frequency band, the geographic area, the time period, and the restrictions on use). The second step is assigning the licenses to particular companies. Although my focus has been on assigning the licenses, since that is what the spectrum auctions are asked to do, the allocation step often is more important. Arguably the greatest economic gains will come from better allocation of spectrum, rather than from improved methods of assigning the spectrum. This is because current spectrum auctions already are highly efficient. In contrast spectrum allocations often are far from efficient.

Determining spectrum allocations involves complex political, engineering, and economic factors. Finding suitable spectrum for new uses is difficult. Often there are incumbent spectrum users. Ideally, one would want market-based tests to determine what spectrum should be auctioned and how it should be structured, but such tests are hard to construct. Political compromises frequently trump good economics. This is especially the case when the competing uses include several constituencies, such as broadcasting, commercial, public safety, and military use.

Finally, the allocation can have a pronounced effect on the success of a particular auction design. For example, having five licenses in the U.K. UMTS auction, rather than four, was critical in stimulating competition (Klemperer, 1998).

9.2. Use care when modifying successful rules

Simultaneous ascending auctions overall have been highly successful in the US and many other countries. Recent FCC auction rules should be a starting point for any government considering spectrum auctions. Modifications to the rules should be considered carefully. The various rules interact in often subtle ways. An apparently innocent change can have disastrous consequences. The interaction between waivers and the declining schedule of minimum opening bids in the Dutch UMTS auction is an example.

Still it is important to recognise that different settings often require a different auction design, as is emphasised in Klemperer (2000). For example, an ascending auction may be inappropriate in situations where competition is weak (Cramton, 1998a).

9.3. Allow discretion in setting auction parameters

The simultaneous ascending auction has a number of parameters (minimum opening bids, minimum bid increments, activity requirements, and rounds per day) that let the government control the pace of the auction. It is fine for

the government to specify guidelines on how it is likely to set parameters, but eliminating all discretion is a bad idea. There are simply too many unknowns going into the auction for the government to specify all parameters in advance. The parameters should be adjusted during the auction to balance the goals of a timely assignment and an efficient assignment. The bidders need time to adjust strategies in light of information revealed in the bidding. Too much haste may lead to bidder error and inefficient assignments. Time also may be needed for bidders to secure additional capital if prices are higher than expected. On the other hand, setting the bid increment too low (e.g., 1 percent) near the end of the auction can result in days of bidding without much progress.

9.4. Reduce the effectiveness of Bidders' revenue-reducing strategies

The information and flexibility available to the bidders in a simultaneous ascending auction is a two-edged sword. Although desirable in reducing bidder uncertainty and promoting efficient license aggregations, the information and flexibility – in certain circumstances– can be used to reduce auction prices. In particular, revenue-reducing strategies may be effective when bidder competition is weak and when bidders already have a sense of who should win what. In this case, the auction is best thought of as a negotiation among the bidders, in which bidders are only able to communicate through their bids. The auction ends when there are no disagreements about who should win what.

As discussed earlier, one common revenue-reducing strategy is demand reduction: the tendency for a bidder to reduce its spectrum demands, knowing that demanding less will tend to reduce spectrum prices. This is a unilateral strategy that is best addressed in the choice of the band plan and geographic scope of the licenses. License structures that make it more difficult for the bidders to split up the spectrum are less vulnerable to demand reduction. For example, offering large nationwide licenses, in which no bidder can win more than one, prevents the bidders from splitting up the spectrum at auction.

The second revenue-reducing strategy is retaliation. This can be thought of as coordinated demand reduction. It is sending another bidder the message that they should stay off your licenses, if they want you to stay off their licenses. Retaliation is especially clear in early auctions where it is possible to use the trailing digits of bids to identify relevant markets. The bidders are effectively able to say things like, "I'll stay out of New York, if you stay out of Los Angeles." This tactic is eliminated by truncating bids to three significant digits, or only allowing bids in integer multiples of the bid increment. However, it is still possible for bidders in certain circumstances to use retaliation to keep prices low. Retaliation is best minimised through careful choice of activity rules, minimum opening bids, and bid increments. An anonymous auction can be used in settings where collusion is likely.

9.5. Use spectrum caps to limit anticompetitive concentration

A spectrum cap is a direct method of limiting the concentration of spectrum for a particular type of service in a particular area. Its advantage is that it is a bright-line test that is easy to enforce, both before and after the auction. In the US, it has played a critical role in ensuring that there are many competitors for mobile wireless services in each market. This competition has led to clear gains for consumers. Its disadvantage is that it is overly simplistic. Spectrum caps cannot take into account the specifics of each situation, and determine whether consumers would be made better or worse off with greater concentration of ownership.

The best policy on spectrum caps is a middle ground, where binding caps are imposed in initial auctions, but then these caps give way once it is believed that vigorous competition has been established. Then individual mergers can be reviewed on a case-by-case basis.

In setting and revising spectrum caps, governments should err on the side of too stringent a cap, since it is much harder to break up a firm than to allow a merger. If concentration is viewed as a potential problem going into an auction, then spectrum caps, rather than case-by-case review, must be used, since only caps can provide an instantaneous determination of what is allowed and what is not. Such a rapid response is essential in a simultaneous ascending auction. Bids must be binding commitments until they are topped. Hence, at every point in the auction, the bidders must know what is allowed and what is not.

Typically, spectrum caps lower auction revenues, but there is one important exception. In situations where incumbent bidders have an advantage, a spectrum cap may actually increase revenues and promote efficiency. In such a situation without a spectrum cap, non-incumbents may be unwilling to participate in the auction, knowing that the incumbents will ultimately win. As a result, in the auction without the cap only the incumbents show up, there is a lack of competition, and the incumbents split the licenses up among themselves at low prices. With the cap, the non-incumbents know that non-incumbents will win licenses, giving them the incentive and ability to secure the needed financing from capital markets. A competitive auction with market prices results. This phenomenon of incumbent bidders getting good deals, because of a lack of non-incumbent competition has been seen in some US auctions, but is most vivid in the Dutch UMTS auction.

9.6. Implement special treatment for designated entities with care

One of the auction objectives that the US Congress gave the FCC is to have a diversity of auction winners. The FCC was instructed to encourage participation by small businesses and women- or minority-owned firms, so called 'designated entities.'

While small and diverse owners may well be a desirable goal for broadcast media with editorial content, the same arguments likely do not apply to mobile wireless communications. Special treatment to designated entities is to some extent premised on the idea that small is beautiful. But what we have learned in the last several years is that there are significant scale economies in wireless communications. Part of the scale economy is the bargaining advantage it creates with equipment suppliers. Another part is scale economies in marketing. But perhaps the largest is the value that consumers place on seamless roaming. As a result, the marketplace has shifted toward nationwide services in most wireless categories. These nationwide services are necessarily billion dollar deals, or tens-of-billions in the case of broadband mobile services. What consumers need is a variety of strong national competitors. In many cases, the small regional players cannot compete. The designated entity rules may simply be setting up the small businesses for failure. This is not desirable, especially given that the FCC's unjust enrichment rules, discussed below, effectively prevent resale to the higher-valued use should failure occur.

On balance, the best policy may be to abandon favours to designated entities, and to use spectrum caps to guarantee new entry where desirable and to prevent over-consolidation of spectrum. An alternative is to offer non-incumbents bidding credits to encourage new entry. My reason for this conclusion has to do with the practical difficulties of effectively implementing favours for designated entities, which I discuss below.

The FCC has used bidding credits, set-asides, and installment payments to encourage the participation and success of designated entities. The idea is that without special treatment, these small businesses would find it difficult to compete with the large incumbents. The favoured treatment can serve to 'level the playing field,' and thereby foster innovation and intensify competition.

Although this is a valid point in theory, and even has some empirical support (Ayres and Cramton, 1996), governments must be cautious when using favours for designated entities. A vivid example is the FCC's only major setback, the C-block broadband PCS auction. (Other disappointing auctions were IVDS and WCS, but none have involved the economic loss seen in the C-block.) The auction failed largely because of overly attractive installment payments (10 percent down and 6-year interest-only at the risk-free 10-year Treasury rate). This encouraged speculative bidding, which led to all the major bidders defaulting and declaring bankruptcy. Even now, years after the auction, much of this C-block spectrum lies unused, tied up in bankruptcy litigation. Instalment payments were a bad idea, because they advantaged the bidders with the most speculative business plans. In addition, installment payments put the FCC in the role of banker, an activity for which the FCC has no advantage. Since the C-block experience, the FCC no longer offers installment payments.

The two other instruments to favour designated entities – set-asides and bidding credits – may be desirable in special situations. The typical situation is one where

the government is attempting to encourage competition in the auction and the post-auction market for wireless services. By levelling the playing field between incumbents and new entrants, competition may be enhanced.

Still, set-asides and bidding credits have serious potential problems. Gauging the right level of set-asides or bidding credits is extremely difficult. Also, it is nearly impossible to target the favour to the desired group. The creation of fronts, carefully constructed to satisfy the rules but circumvent their intent, has been a constant problem in the FCC auctions.

One general rule, whether using set-asides or bidding credits, is that it is best for incumbents and non-incumbents to compete in the same auction. Then if competition among non-incumbents is sufficiently robust, the non-incumbents will be able to spill over to the licenses that incumbent bidders can bid on. This spillover increases competition, and hence revenues in the auction.

Another problem with favours for designated entities is their impact on the resale of spectrum. The FCC auction rules prohibit resale to a non-designated entity for a period of time, and include an 'unjust enrichment' provision that requires that the FCC be paid back the bidding credit plus interest. The reality has been that the bidding credits are often bid away by competition among designated entities. Indeed, even after accounting for the value of the installment payments and the bidding credits, the C auction resulted in prices that were well above what the large firms paid in the AB auction. Given these facts, it is difficult to understand why the small firms are required to pay a huge 'unjust enrichment' penalty, when there is no unjust enrichment. As it stands, the penalty is so large that it is often an insurmountable barrier to trade.

Perhaps the most serious problem with favours to designated entities is that they greatly complicate the auction process. Too often the rules for designated entities become a central issue in establishing the auction procedures. These rules are complex. They are difficult to write, difficult to enforce, and difficult to defend. The absolute worst outcome in a spectrum auction is having the licenses tied up in litigation. Until the litigation is resolved the building of communication services cannot begin. Even the risk of litigation can have a disastrous effect on auction participation, and hence revenues.

9.7. Implementing an effective auction; time and difficult tradeoffs

A second FCC auction disappointment was the Wireless Communication Services (WCS) auction, held in April 1997. Revenues in this auction were a tiny fraction of what they might have been. The main problem was the stringent out-of-band emission limit. Equipment manufacturers warned that this would threaten the commercial viability of this spectrum. The low prices at auction and the absence today of activity in this band appears to confirm that the equipment manufacturers were right. At the time of the decision, the FCC was facing a difficult trade-off between the rights of prior winners of neighbouring licenses and the WCS

use. Such decisions are always difficult, but the FCC was under intense time pressure to meet the timetable that Congress set for the auction. This aggressive timetable may well have led the FCC to make a too-hasty decision on interference rules, which damaged the value of this spectrum. Congress's desire for receiving revenues according to its fiscal calendar may have resulted in substantially reduced auction revenues.

9.8. Facilitate efficient clearing when auctioning encumbered spectrum

An issue of increasing importance is the auctioning of encumbered spectrum. Many of the FCC auctions are for overlay licenses (the PCS and 700 MHz auctions are examples). An overlay license is for the portions of the band that are not occupied by incumbent licensees. Effectively the FCC is auctioning Swiss cheese, where the holes are incumbent licensees with particular rights. These incumbents must either be cleared or worked-around in order for the new entrant to provide a service. Negotiations between the new entrant and the incumbent are often difficult due to holdout by the incumbent. A second problem occurs when multiple new entrants benefit from the clearing of a single incumbent; then each new entrant can hope to free ride on the clearing done by others. These problems often prevent or delay the efficient clearing of the spectrum.

The government can play an important role in adopting rules that promote the efficient clearing of the spectrum by structuring the rules of negotiation appropriately (Cramton et al., 1998). The broadband PCS rule-making is a good example. The FCC adopted rules that went a long way in minimising the holdout and free-rider problems that undermine efficient clearing. New licensees were given the right to move an incumbent after a period of time, compensating the incumbent for its relocation costs. This rule minimises the need for costly negotiations.

The upcoming 700 MHz auction is a more difficult case. This spectrum currently is used for analogue television channels 60–69. The spectrum cannot be used for new services until the analogue broadcasters terminate over-the-air broadcasting on these channels. Because of the political power of broadcasters, the FCC was unwilling to adopt PCS-type rules that promote efficient clearing. As a result, clearing will be much more costly and inefficient.

9.9. Promote market-based tests in spectrum management

Spectrum auctions are a critical step in the march toward market-based spectrum management. Governments should continue on this path. Flexibility should be the norm, not constraints. Constraints should appear only when those constraints help foster a more competitive environment by adding essential structure. (A good example of the benefits of structure is the success of GSM in Europe.) Specifically, governments should:

- Allow service flexibility
- Allow technical flexibility
- Set initial interference rules, but allow trading
- Set initial geographic and bandwidth scope to an *ex ante* view of how spectrum will be used, but allow spectrum partitioning and geographic disaggregation
- Eliminate buildout requirements
- Allow transfers of licenses

Many of the current restrictions are holdovers from the days of comparative hearings. These needless regulations should be eliminated. The rate of technological change is now so great that attempting to craft specific regulations as was done in the past is hopeless and destructive. Rather, governments should focus on broad principles that encourage competition. Legislatures especially should refrain from the micro-management of spectrum policy. The complex economic and engineering tradeoffs are much better left to a specialised agency.

10. Conclusion

Spectrum auctions represent a huge advance in the way governments allocate scare resources. Despite some early bumps, spectrum auctions largely have succeeded in putting spectrum in the hands of those best able to use it. Economists have made (and continue to make) valuable contributions to spectrum auction design. The simultaneous ascending auction, pioneered by the FCC, has been remarkably successfully in the US and other countries. Its simple process of price discovery promotes efficiency, especially in competitive auctions in which bidders do not have large and varied license synergies. The recent introduction of package bidding to this successful format promises to extend the set of environments where the simultaneous ascending auction performs well.

In most cases, the greatest room for improvement lies not in the auction design, but in the allocation process. Spectrum allocations often are the result of political forces that ignore the underlying economics. Governments would do well to let competitive market forces, not political lobbying, determine spectrum use. The auctioning of spectrum is a major step in the right direction.

References

Ausubel, L. M., 1997, An efficient ascending-bid auction for multiple objects, Working Paper, University of Maryland.

Ausubel, L. M. and P. Cramton, 1999, The optimality of being efficient, Working Paper, University of Maryland.

Ausubel, L. M. and P. Cramton, 2001, The 3G European auctions, Working Paper, University of Maryland.

Ausubel, L. M. and P. Cramton, 1996, Demand reduction and inefficiency in multi-unit auctions, Working Paper, University of Maryland.
Ausubel, L. M., P. Cramton, R. P. McAfee and J. McMillan, 1997, Synergies in wireless telephony: Evidence from the broadband PCS auctions, Journal of Economics and Management Strategy, 6, 497–527.
Ausubel, L. M. and J. A. Schwartz, 1999, The ascending auction paradox, Working Paper, University of Maryland.
Ayres, I. and P. Cramton, 1996, Deficit reduction through diversity: A case study of how affirmative action at the FCC increased auction competition, Stanford Law Review, 48, 761–815.
Ballard, C. L., J. B. Shoven and J. Whalley, 1985, General equilibrium computations of the marginal welfare costs of taxes in the United States, American Economic Review, 75, 128–138.
Banks, J. S., J. O. Ledyard and D. P. Porter, 1989, Allocating uncertain and unresponsive resources: An experimental approach, RAND Journal of Economics, 20, 1–22.
Brusco, S., and G. Lopomo, 1999, Collusion via signalling in open ascending auctions with multiple objects and complementarities, Working Paper, New York University.
Bykowsky, M. M., R. J. Cull and J. O. Ledyard, 2000, Mutually destructive bidding: The FCC auction design problem, Journal of Regulatory Economics, 17, 205–228.
Chakravorti, B., W. W. Sharkey, Y. Spiegel and S. Wilkie, 1995, Auctioning the airwaves: The contest for broadband PCS spectrum, Journal of Economics and Management Strategy, 4, 345–373.
Coase, R. H., 1959, The Federal Communications Commission, Journal of Law and Economics, 2, 1–40.
Cramton, P., 1995, Money out of thin air: The nationwide narrowband PCS auction, Journal of Economics and Management Strategy, 4, 267–343.
Cramton, P., 1997, The FCC spectrum auctions: An early assessment, Journal of Economics and Management Strategy, 6, 431–495.
Cramton, P., 1998a, Ascending auctions, European Economic Review, 42, 745–756.
Cramton, P., 1998b, The efficiency of the FCC spectrum auctions, Journal of Law and Economics, 41, 727–736.
Cramton, P., R. Gibbons and P. Klemperer, 1987, Dissolving a partnership efficiently, Econometrica, 55, 615–632.
Cramton, P., E. Kwerel and J. Williams, 1998, Efficient relocation of spectrum incumbents, Journal of Law and Economics, 41, 647–675.
Cramton, P. and J. Schwartz, 1999, Collusive bidding in the FCC spectrum auctions, Working Paper, University of Maryland.
Cramton, P., and J. Schwartz, 2000, Collusive bidding: Lessons from the FCC spectrum auctions, Journal of Regulatory Economics, 17, 229–252.
Engelbrecht-Wiggans, R. and C. M. Kahn, 1995a, Multi-unit auctions with uniform prices, Working Paper, University of Illinois.
Hazlett, T. W., 1998, Assigning property rights to radio spectrum users: Why did FCC license auctions take 67 years? Journal of Law and Economics, 41, 529–576.
Jehiel, P. and B. Moldovanu, 2000, A critique of the planned rules for the German UMTS/IMT-2000 license auction, Working Paper, University of Bonn.
Kagel, J. H., and A. E. Roth, 1995, The handbook of experimental economics, Princeton, NJ: Princeton University Press.
Klemperer, P., 1998, Almost common value auctions: The 'Wallet Game' and its applications to takeover battles and airwaves auctions, European Economic Review, 42, 757–769.
Klemperer, P., 2000, What really matters in auction design, Working Paper, Oxford University.
Krishna, V. and R. Rosenthal, 1997, Simultaneous auctions with synergies, Games and Economic Behaviour, 17, 1–31.
Maskin, E. and J. Riley, 1984, Optimal auctions with risk averse buyers, Econometrica, 52, 1473–1518.
Maskin, E. and J. Riley, 1996, Asymmetric auctions, Working Paper, UCLA.

Matthews, S. A., 1983, Selling to risk averse buyers with unobservable tastes, Journal of Economic Theory, 30, 370–400.
McAfee, R. P. and J. McMillan, 1996, Analysing the airwaves auction, Journal of Economic Perspectives, 10, 159–176.
McMillan, J., 1994, Selling spectrum rights, Journal of Economic Perspectives, 8, 145–162.
McMillan, J., 1995, Why auction the spectrum? Telecommunications Policy, 19, 191–199.
Milgrom, P., 1987, Auction theory, in T. Bewley, ed., Advances in Economic Theory - Fifth World Congress, Cambridge; Cambridge University Press.
Milgrom, P., 2000, Putting auction theory to work: The simultaneous ascending auction, Journal of Political Economy, 108, 245–272.
Milgrom, P. and R. J. Weber, 1982, A theory of auctions and competitive bidding, Econometrica, 50, 1089–1122.
Moreton, P. S. and P. T. Spiller, 1998, What's in the air? Interlicence synergies and their impact on the FCC's broadband PCS license auctions, Journal of Law and Economics, 41, 677–716.
Perry, M. and P. J. Reny, 1999, An ex-post efficient auction, Working Paper, Penn State University.
Rosenthal, R. W. and R. Wang, 1997, Simultaneous auctions with synergies and common values, Games and Economic Behaviour, 17, 32–55.
Rothkopf, M. H., A. Pekec and R. M. Harstad, 1998, Computationally manageable combinatorial auctions, Management Science, 44, 1131–1147.
Vickrey, W., 1961, Counterspeculation, auctions, and competitive sealed tenders, Journal of Finance, 16, 8–37.

This Page Intentionally Left Blank

Chapter 15

LOCAL NETWORK COMPETITION

GLENN A WOROCH*

University of California, Berkeley

Contents

* I have received helpful comments on earlier drafts from Mark Armstrong, Bob Crandall, David Gabel and Lester Taylor. Ellen Burton of the FCC and Amy Friedlander of CNRI supplied useful empirical and historical information. I am especially grateful to Ingo Vogelsang for his encouragement, guidance, and inexhaustible patience throughout this project.

Handbook of Telecommunications Economics, Volume 1, Edited by M.E. Cave et al.

1. Introduction

1.1. Scope and objectives of this chapter

This chapter surveys the economic analysis of competition in markets for local telecommunications services[1]. Its main objective is to understand patterns of competition in these markets and evaluate its benefits and costs against the alternative forms of industrial organisation. While regulation is not a principal focus, policies governing rates and investment of providers – incumbents and entrants alike – can greatly affect the extent of competition. More recently, the opening of incumbent networks and the unbundling of network services for sale to competitors has become the preferred means to move toward competition in these markets. Technology has the potential to make all this policy irrelevant by sweeping in a new generation of competitors offering innovative services and driving out incumbent providers. Good examples of such a potential is how fixed and mobile wireless technologies could replace traditional wireline services, and how packetised voice and data can run on many alternative media, not just the traditional public switched telephone network (PSTN). At the same time, new technologies could have the effect of solidifying the dominance of incumbent providers.

The term 'local' in the chapter title deserves some explanation. As usual, it has a spatial meaning, but one that is being redefined all the time by changes in technology and public policy. During the very earliest days of the industry, the geographic market ended at the city limits lacking the technologies to overcome the signal attenuation experienced by long distance transmission. Today, as always, much of demand for communication reflects the local nature of social relationships, and so this chapter will include the provision of switched voice services within an urban area. The meaning of local becomes more challenging to define, however, when facilities that provide these services also connect users with individuals and machines located far away. The recent debates over the meaning of local when facilities carry Internet traffic illustrates the difficulty of arriving at a sharp delineation of these markets.

While much demand for communication may still be spatially local, the scope of supply may be far less limited. It may be efficient for a single provider to serve many local areas. Furthermore, it may be technically efficient and strategically advisable for a local service provider to offer customers "non-local" services as well, such as long distance and Internet access. What distinguishes suppliers to these markets is that they provide originating and terminating legs of a communication link, whether it carries voice or data, or whether it is over wireline or wireless facilities.

[1] Other recent surveys of local telephone competition include Baumol and Sidak (1994), Vogelsang and Mitchell (1997), Vogelsang and Woroch (1998) and Laffont and Tirole (2000).

In the past, telecommunications has been synonymous with voice communications. Increasingly that term has come to include one-way image and video transmissions and interactive data services. We will use the broader interpretation here in the context of local markets. Many services often associated with the telecommunications sector will be excluded however, including video and audio broadcasting (whether over the air or on cable), Internet access, services and content (though we would include dialup access over local loop), and long distance and international service (except to the extent local networks provide origination and termination for these services). We will also exclude the creation and modification of content, the manufacturing of communications equipment and the development of network software.

The reader will soon see that many of the examples found in this chapter are drawn from experiences in the United States, both current and historical. This is limiting to the extent that different paths were followed with different results outside the U.S. At the moment, the range of experiences is particularly broad, as countries experiment with a wide array of approaches to local competition, invariably from the starting point of a state owned monopoly. I will allude to a few of these experiments, keeping in mind that they began well after the opening of markets to competition in the U.S.

1.2. Patterns and themes

The study of local network competition – especially under the world's new institutional and regulatory structures – remains in its infancy. In so many cases it is too soon to register the full impact of new policies toward these industries. And as usual, theory is way ahead of empirical testing. Nevertheless, there are a few distinct patterns that emerge from the record, drawing as well on early history of the industry and the experiences outside the U.S.

To begin with, no inexorable, inherent tendency toward monopoly or toward competition can be discerned from the history of this industry. The past century witnessed several major transformations, first from unregulated monopoly to fierce competition, and then to regulated monopoly, and most recently to (de)regulated competition. Regulation and technological change played key roles in each case – in addition to luck and serendipity. The first episode of competition began when the end of the Bell patent monopoly threw open the doors to local markets to new entrants. The duplication and waste attributed to this period fed public opinion that competition does not work in this industry, and eventually led to creation of a monopoly franchise reined in by an elaborate regulatory institution in the U.S. and state ownership elsewhere.

Nowadays the view is that regulation does not work and competition is the solution (and to a lesser extent that technology has advanced to the point where competition is viable). Deep dissatisfaction with administrative regulation and a faith in the discipline of the market place, along with help from an endless stream of

technological innovations, resulted in rebuilding the regulatory infrastructure, to aid competitors of all kinds and to free up incumbents. Ironically, over the near term, government intervention has expanded to guide this transition to competition.

For much of its history, the local telephone industry was thought to be naturally prone to monopoly as a consequence of massive scale and scope economies in provision of services over wireline networks. These economies are still present today but now there are other technologies that have cost characteristics that may support competition. What remains unchanged, however, is the fact that incumbent suppliers enjoy strategic advantages that tend to fortify any initial advantage they may acquire. Sunk facilities have always been a means to gain a first mover advantage, but now it is recognised that such advantages stem from several other sources. In particular, 'network effects' of certain services and user switching costs make creating a competing network difficult or impossible. New technologies (such as instant messaging) may be no less susceptible to dominance than more traditional networks. These technologies may be capable of supporting more firms but first mover advantages may make it exceedingly difficult for them to amass a customer base necessary to cover their entry costs.

Dramatic shifts over time in the consensus regarding the relative merits of competition and monopoly in the local network shake one's confidence in the wisdom of the prevailing view. Only with great humility can anyone claim that competition is desirable and sustainable given the likelihood of technological change and untapped opportunities for institutional innovation.

2. Local network competition in historical perspective

2.1. Local competition in one city, a century apart

As 1894 began, residents of New York City were served by either of two telephone companies: Metropolitan Telephone & Telegraph Co. and New York & New Jersey Telephone Co[2]. For several years, each company had operated under a license to use Alexander Graham Bell's basic telephone patents. But now those patents had expired and, free to exploit these technologies, several companies entered this highly lucrative market with its large population and rapidly growing business community. Adding to the attraction, New York telephone customers were widely dissatisfied with Bell rates and service[3]. Indeed, the average annual charge for local service in the city was $253 in 1915 compared to a nation-wide Bell-company average of $30.93[4].

[2] Wilcox (1910, p. 258).
[3] Gabel (1994, p. 561).
[4] U.S. House of Representatives (1915, p. 24). Note that the next three largest cities at that time, Chicago, Philadelphia and St. Louis had average annual rates of $84, $90 and $78, respectively.

At the end of the Bell monopoly in 1894, Mercantile Electric Co. made plans to build an exchange for bankers and brokers, and New York & Eastern Telephone Co. applied to provide service in Brooklyn and Manhattan[5]. Over the next several years, People's Telephone Co., Atlantic Telephone Co., and New York Electric Lines would each make separate bids for some part of the New York City phone market[6]. Each of these entrants would meet with opposition from state and local regulators and from the Bell interests. An 1885 New York state law required all phone lines to be buried underground in the city streets[7]. A monopoly over the underground structures was awarded to the Empire City Subway Co., Ltd. as compensation for undertaking this risky investment[8]. Importantly, a major owner of Empire City Subway was the Bell System itself. As a consequence, entrants into this market were forced to secure essential rights of way from a direct competitor, and it was no surprise when they were told there was no free space and/or charged high access fees while the Bell companies did not directly pay anything for the same rights. Municipal authorities demanded sizeable franchise fees and required competitors to achieve interconnection with an overwhelming percentage of the long distance providers in a short period of time. Unconvinced of the benefits of local telephone competition, New York City municipal authorities would repeatedly deny requests to enter this market[9].

In the end no independent telephone company would break into the New York City market. On the contrary, the many different operating companies in the city and upstate consolidated into a single operating company, New York Telephone (NYT). Competition nevertheless left its mark, with average monthly charges falling by more than a half before it was over[10].

Some ninety years later, phone competition would again break out in New York City, but this time it would be more methodical and less visible. In 1982, Merrill Lynch and Western Union formed a joint venture to build a 'satellite park' on Staten Island, one of the five boroughs of New York. This investment was a response to the capacity crunch in the region exacerbated by the explosive growth in the financial services industry, and the companies' desire to have access to a highly reliable network. In a couple of years the project, now privately owned under the name "Teleport," constructed an optical fibre link to Manhattan where it could gather traffic to put out over the satellite network. Eventually, the fibre network would extend throughout lower Manhattan reaching some of the world's most communications-intensive customers. The fibre would be strung under the streets of New York using space owned by none other than the Empire City

[5] Mueller (1997, p. 55).
[6] See Gabel (1994) and Wilcox (1910).
[7] Merchants Association of New York (1905, p.15).
[8] Wilcox (1910, p. 260).
[9] The Brooklyn city council franchised an independent three times during this period only to be vetoed by the city manager on each occasion. Mueller (1997, p. 62)
[10] Mueller (1997, p. 66).

Subway Co., the same company that obstructed competitors at the end of the previous century.

Fibre optic transmission was a new communications technology but one that had been introduced to the New York market earlier. In 1979, New York Telephone deployed fibre in its interoffice network in Brooklyn[11]. A further difference was that Teleport was laying fibre right to customer's buildings. Also, it built a network that had an extraordinary high level of reliability greatly desired by the financial community and others in the New York area.

In 1986 New Jersey Bell – no longer a part of AT&T – agreed to provide collocation to Teleport's network in or near Bell's central offices in the Newark and Jersey City region. This arrangement allowed Teleport to tap into traffic gathered by New Jersey Bell's network without building out facilities to all the customers. When Teleport sought access to New York Telephone's central offices in New York City, it received a very different reception. NYT claimed that free space in its central offices was scarce and insisted on charging Teleport its retail tariff rates for originating and terminating traffic. After much negotiation, the New York Public Service Commission (NYPSC) ordered NYT to provide Teleport and other alternative access providers with "comparably efficient interconnection" for intrastate private line and dedicated access services[12]. It also required NYT to unbundle its "links and ports" at its switches to enable companies like Teleport to offer its customers local services directly comparable to what NYT offered.

Teleport went on to build fibre ring networks in over 50 cities in North America, Europe and Asia. Ironically, Teleport was purchased in 1998 by AT&T for $11.3 billion and now represents the core of its local business services division.

The two competitive episodes that occurred in New York City are interesting to compare. In both cases a dominant incumbent was exposed to facilities-based competition from *de novo* entrants. Success of new entrants turned on their access to rights of way – especially underground conduit – and to collocation and interconnection with the incumbent network. The earlier competitors failed to achieve viability as facilities-based carriers whereas the more recent competitive carriers made huge incursions into Bell market share, especially among large corporate accounts[13]. In the earlier era, business customers rallied against competition, demanding a single franchise provider to avoid the expense of a dual system.

Big changes had taken place over the course of a hundred years, however. In the nineteenth century incumbents and entrants competed for switched local

[11] Mikolas (1990).

[12] Much earlier in 1985, the New York PSC authorised Teleport to compete with NYT in New York.

[13] For instance, Deutsche Bank Alex. Brown (2001) estimate that competitors supply roughly half of all private lines in New York City.

service (though long distance interconnection played a role). In the twentieth century version, they vied instead for long distance access and private lines. Most importantly, earlier competition failed to take hold while by all accounts competition in both residential and business services markets in today's New York City is vibrant and unlikely to return to monopoly any time in the foreseeable future[14].

Some may argue that the New York City experience does not transfer over to other markets. Indeed, no-where in the U.S. is population density and volume of telecommunications traffic greater than in New York City. Nevertheless events that played out in New York were repeated in large urban areas across the country in the late 1980s and throughout the 1990s – including many of the same regulatory struggles for entry into those markets.

2.2. U.S. experience with local competition and monopoly

2.2.1. The Bell patent monopoly: 1876–1894

As is so well known, in 1876 a teacher of the deaf, Alexander Graham Bell, filed for patents on the telephone transmitter that he called "An Improvement for Telegraphy." Also well known is the fact that only hours later Elisha Gray filed a "caveat" with his intent to file an application with the Patent Office for an invention that also transmitted sound over wires. Western Union, the telegraph behemoth, acquired Gray's device a year later and hired Thomas Edison to perfect the talking telephone. Western Union, the target of a patent infringement suit by the Bell interests, would agree in 1878 to withdraw from the local phone business in exchange for a 20 percent royalty on revenue received by Bell's National Bell Co. through to the expiration of its basic patents.

Bell licensed operating companies to use his telephone technology in mutually exclusive geographic regions. This allowed Bell to deploy the technology quickly without the huge financial burden of building out the networks. In exchange, Bell would receive license fees for the patented telephone technologies usually calculated on the number of instruments rented to customers. In time, the company would begin to take an equity stake in the operating companies, giving it control over the pricing, investment and other strategic decisions, including interconnection with other local telephone companies[15].

During the patent monopoly period, Bell concentrated efforts on selected markets. Licensees initially built systems in large cities in New England and along the Atlantic coast. When the period came to a close, Bell networks were

[14] In approving Bell Atlantic's application to enter the long distance market in New York State, the FCC not only confirmed that the company had met the 14-point checklist for opening its local markets to competition, but concluded that the openness was irreversible. See Federal Communications Commission (1999, pp. 429–443).

[15] Brock (1994, p. 63).

concentrated in the largest population centres[16] and focused on business customers[17]. Bell devoted relatively little effort during these years to developing the local telephone service except to steadily increase its equity stake in its operating companies. Instead it continued to build up its patent arsenal (it was granted an additional 900 improvement patents) and aggressively defended its intellectual property (filing over 600 patent infringement suits during 1877–1893)[18]. It also invested in long distance infrastructure correctly foreseeing that intercity service would be an extremely valuable complement to local service.

2.2.2. Early competitive era: 1894–1907

Patents over the two basic telephone instruments – the transmitter and the receiver – expired in 1893 and 1894, respectively, ushering in a competitive landrush as any operator could then freely use Bell's technology. By 1894, there were several dozen independent telephone companies, all of which were providing local service. Less than ten years later, no fewer than 1,074 commercial, independent phone companies were operating in the U.S[19].

Although independents appeared in the smaller cities and rural areas that did not interest Bell, they also directly attacked Bell's urban turf as the New York City story illustrates. A product of this head-to-head competition was the creation of "dual systems" in which two (or more) facilities-based local telephone companies served the same areas of the same cities. The incidence of dual systems was remarkably high. By 1902, less than 10 years after competition was unleashed, of the 1,051 cities with population of 4,000 or more, 1,002 had telephone service, and 451 of these (or 45.1 percent) had two or more local providers[20]. By 1907, 59 percent of cities and towns with population exceeding 5,000 had dual exchanges[21]. It is estimated that 8–13 percent of subscribers in dual system cities took service from more than one phone company[22].

Dual systems necessarily resulted in duplicate investment, not only in network facilities but also with the multiple handsets, phone numbers and directories that were maintained by homes and businesses. Businesses were especially adverse to the dual system because they saw it as a competitive necessity to subscribe to all

[16] By 1894, only 52 towns had phone service of the more than 7,000 towns with a population exceeding 10,000. Also, 346 of largest cities having 27 percent of the population had 83 percent of the phones. Mueller (1997, pp. 40–41).

[17] As of 1894 90 *percent* of the 240,000 lines were rented to businesses whereas the penetration rate among U.S. households was 1 in 225 (Mueller, 1997, p. 40).

[18] Barnett and Carroll (1993, p. 101).

[19] Bornholz and Evans (1983, pp. 11–12). Several thousand more independents were formed, and went out of business, if municipal and rural systems are included.

[20] Gabel (1969, p. 345).

[21] Mueller (1997, p. 111). Just a half dozen years later that figure would drop to 33 *percent*.

[22] Bornholz and Evans (1983, p. 18).

local networks. They were not persuaded that the lower prices that derived from competition compensated them for the additional costs.

AT&T responded aggressively to the independents. In addition to defending its patents, the company refused to supply independents with switching and transmission equipment from its manufacturing arm, Western Electric. In each market where AT&T met with competition from independents, the company slashed prices and refused to interconnect with the independent's network.

The intense competition for local service had a dramatic impact on the industry. Average amounts paid for phone line rental fell (nominally) from $5.74 per month in 1893 to $1.45 in 1898, a fall of 75 percent[23]. Unquestionably, diffusion of telephone service in the U.S. was accelerated by the price cuts. A total of 258,455 lines in 1893 more than doubled to 562,423 lines five years later[24]. Over this same period, the number of phone per 100 population would more than double from 3.9 to 9.2. Not surprisingly the competition took its toll on profits of AT&T and the independents alike. Whereas AT&T enjoyed a 46 percent profit rate during the patent monopoly period, its rate fell to 8 percent in the competitive period[25].

While ownership of independent telephone companies was unconcentrated, their overall strategies were co-ordinated to some extent through trade associations. Like AT&T, independents refused to interconnect their local networks. The independents also formed a long distance network to serve their local networks, but symmetrically with Bell, they resisted connecting with AT&T's Long Lines Division.

2.2.3. Regulated monopoly: 1907–1956

1907 was a watershed year for the early telephone industry. In that year independent telephone companies reached their peak by securing 51 percent of all phones, or 3.1 million out of a total of 6.1 million[26]. In that same year the first state public utility commissions with powers to regulate local telephone service were formed in Wisconsin and New York. And no less significant, Theodore Vail was made chief executive officer of AT&T.

Upon taking charge at AT&T, Vail quickly made major corrections to the company's strategic direction[27]. While he called for an end to the aggressive price wars in markets where the company faced local independents, Vail slashed toll prices by around two-thirds where AT&T competed with independent long distance carriers. He also accelerated the acquisition of independent local

[23] Stehman (1925).

[24] Stehman (1925) op. cit. which also indicates very slow growth during the monopoly period. For instance, between 1888 and 1893, lines per 100 population only grew from 3.2 to 3.9.

[25] Brock (1981, p.117).

[26] Bornholz and Evans (1983, p. 13).

[27] Barnett and Carroll (1993, p. 112).

companies and equipment manufacturers, and instructed the company's Long Lines Division not to interconnect with independent companies who served the same markets as Bell operating companies (and also some markets where Bell was not present).

Vail, along with many others, had embraced the contemporary notion of 'natural monopoly' and adapted it to the telephone industry. This early version of the concept held that a single firm could best serve the public judged by the quality, reliability and coverage of its service. Vail did not argue that a monopoly delivered service at least cost, but he did claim that competition led to 'unnecessary duplication' characteristic of dual systems[28]. He made clear his willingness to accept reasonable government regulation in exchange for protection from competition, an offer the government eventually accepted.

The country was already on its way to creating the telecom regulatory institutions when, in 1910, the Mann-Elkins Act empowered the Interstate Commerce Commission (ICC) to control AT&T's rates and accounting methods. By 1920, 45 of 48 states had given their public utility commissions the power to regulate local telephone service[29].

In 1913, responding to an antitrust investigation, AT&T came to agreement with the U.S. Department of Justice. In the 'Kingsbury Commitment,' Bell promised to halt acquisition of competing independents and to interconnect with non-competing independents (provided they satisfied its technical requirements). The company also agreed to divest itself of Western Union Telegraph Co.

It is during this period that the so-called ''Bell System'' was formed out of 22 wholly owned operating companies, plus Western Electric, Bell Laboratories and the Long Lines Division. Despite the Kingsbury Commitment, AT&T continued to acquire independent phone companies, though this was balanced against shedding of properties outside of large population centres, leaving smaller towns and rural areas to independents. By this time, Apartheid of telephone carriers was complete, with markets divided between Bell and independent companies. Independents' share had fallen to 21 percent with 100 percent connected to AT&T's long distance network by 1934, the year that the landmark Telecommunications Act passed. This Act crystallised the regulatory superstructure that had been taking shape for many years. It created the Federal Communications Commission (FCC) with powers over interstate telecommunications services, a jurisdiction that included local facilities used to provide access to these services. At both the state and federal levels, local service was subject to some form of rate base-rate of return regulation (RB-RORR). This quasi-judicial procedure set rates so as to ensure a 'reasonable' return on invested capital, and controlled which investments were allowed a return.

[28] AT&T's 1910 annual report.
[29] Stone (1997).

2.2.4. Early transition to competition: 1956–1984

By the end of World War II, the Bell System dominated the local telephone industry in the U.S. with the Bell operating companies (BOCs) accounting for over 90 percent of all local lines. This success placed the company in the cross hairs of antitrust authorities, and in 1949 the Department of Justice launched another investigation of AT&T, focusing this time on its ownership of Western Electric. Eventually the two parties would sign a consent decree in 1956 barring the company from providing non-telephone services and building non-telephone equipment, and forcing it to license its patents at reasonable royalties.

Perhaps more significant in its implications for local competition was a Court of Appeals decision in the Hush-a-Phone case that same year[30]. In that decision, following years of FCC flip flopping on the case, the Court overturned two previous FCC rulings that concluded that the Hush-a-Phone device, a metal attachment to the handset that enhanced privacy of phone conversations in a crowded room, jeopardised the integrity of the public network. Instead the Court decided such devices were legitimate provided they were "privately beneficial without being publicly harmful." Here was the first major crack in the monolithic Bell network[31].

Another major hole was created in 1959 when the FCC rendered its 'Above 890' decision[32]. It concluded that there was enough spectrum to allow users to build private microwave networks for voice transmission. Not only did this lay a foundation for competitive long distance companies such as the nascent Microwave Communications, Inc., but also microwave-based bypass providers that appeared on the urban scene in the 1980s. With time, these high frequency bands would be the basis for fixed wireless access methods that are being deployed today.

The next major opening of local markets came in 1968 when the FCC ruled that another device that interconnected the phone network with a private radio system was allowed. In its 'Carterfone' ruling, the FCC articulated some of the first principles of a federal interconnection policy[33]. It expanded allowable competition beyond the Hush-a-Phone decision which involved a network attachment, to include an interconnection device.

A characteristic of this period was judicial leadership in supporting competition into both equipment and long distance service, with widespread

[30] Hush-a-Phone Corp. v. U.S., 238 F.2d 266 (D.C. Circuit, 1956).

[31] It was not until 1980, that the FCC, as part of its Computer II Inquiry, would move to fully decontrol customer premise equipment. In that decision, the Commission would allow AT&T and GTE to sell equipment through structurally separate subsidiaries.

[32] Allocation of Frequencies in Bands Above 890 Mc, 27 F.C.C. 359 (1959).

[33] Carterfone, 14 F.C.C. 2d 571 (1968).

reluctance among state and federal regulators. Once again in 1974, the U.S. Department of Justice, goaded by new entrants into telephone markets, launched an antitrust investigation of possible abuse of monopoly power in violation of the Sherman Antitrust Act. Another protracted investigation and trial, this effort would again result in a consent decree between the government and AT&T. This time, the terms would completely transform the local telephone industry.

It would be a mistake to focus entirely on judicial and regulatory explanations for the competition during the post-war period. In fact, technological developments fuelled local competition. Primitive as they may seem by today's standards, the Hush-a-Phone and Carterfone devices were innovations that provided vertical services to the public network; earlier we mentioned microwave transmission as a bypass technology that appeared in this time frame. The microelectronics revolution would enable innovations in switching and transmission technology that again expanded opportunities for competition in local markets. Electronic stored program control switches greatly accelerated connections, and allowed carriers to add new features to basic service by rewriting a software program. These developments also produced the 'private branch exchanges' (PBX). These allowed business customers to displace switching that otherwise would be provided by their local carrier.

Optical fibre technology, first used for communication in the late 1970s, would revolutionise transmission. Initially, optical fibre was deployed in long distance networks to replace microwave transmission. Soon after, local carriers began to replace their interoffice trunks with fibre. And as described earlier, fibre was the killer technology supporting the entry of competitive access providers in high-capacity local access services.

Developments in the wireless technology during this period laid the foundation for what might become the greatest threat to wireline local service. Bell Labs developed the first analogue cellular telephone technology, Advanced Mobile Phone System (AMPS), back in 1947 but did not gain approval to deploy it commercially until 1982[34]. As we will discuss below, when cellular telephone was first rolled out in Baltimore and Chicago in 1983, it did not offer much competition for wireline service. Its price was high, the signal quality and coverage were poor, and the heavy, cumbersome phones were bolted into automobiles. Furthermore, the FCC licensed two carriers for each urban and rural market with one license reserved for the local wireline carrier. The cellular duopoly did not engender much wireless competition, but that changed considerably in 1995 when the FCC licensed up to five additional carriers of Personal Communications Services (PCS) for those very same markets.

[34] See the chapter by Hausman in this *Handbook* for a description of the history of commercialisation of the AMPS technology in the U.S.

2.2.5. Competition by divestiture and deregulation: 1984-present

Ending an 8-year investigation and trial, AT&T and the Department of Justice (DOJ) signed a consent decree on January 1, 1982. Called the Modification of Final Judgment (MFJ) because it amended the 1956 consent decree the parties had signed over 25 years earlier, this historic agreement called for a divestiture of AT&T, including the severing of the local operating companies from the rest of the company. The operating companies were grouped into seven Regional Bell operating companies (RBOCs) that were geographically quarantined to 162 'local access and transport areas' or LATAs[35]. These were judicial boundaries not necessarily reflecting geography of economic markets.

The RBOCs were allowed, if not encouraged, to enter each other's territories to provide local service. In addition, the MFJ contained no explicit wording that kept AT&T out of these areas, and hence offering its own local service. Other line of business restrictions (LOBs) banned the RBOCs from equipment manufacture and the provision of long distance and enhanced services (except when approved by the Court overseeing the MFJ). Consequently, if they attempted to enter local markets outside their region, they could not bundle equipment and long distance service with local service – a strategy that helped AT&T gain its dominant share in early years of the industry.

An important aspect of the AT&T Divestiture – supplemented by a series of FCC orders – was the nurturing of long distance competition emerging at the time. Indirectly, this long distance competition advanced local competition by exerting pressure on access service markets: thinner margins in long distance drove interexchange carriers and their largest customers to seek cheaper alternatives to RBOCs' access fees.

Competitive access providers, or CAPs, filled this need. Companies like Teleport in New York built high capacity transport networks to interconnect interexchange carriers and to deliver toll access to large business customers. CAPs with their new technology alone could not bring competition to the local exchange, however; regulatory reform was needed to support multiple providers where the incumbents had enjoyed *de facto* franchise monopolies. Local competition initiatives, such as the NYPSC interconnection decision in New York City, exemplified the innovative experiments that were taking place at the state level. Typically, these proceedings were initiated by the entrant phone companies and facilitated by state regulators who eventually mediated an agreement among the companies.

Another good example, that again took place in New York State, was the restructuring plan proposed by Rochester Telephone, called "The Open Market Plan"[36]. After modification by the NYPSC, Rochester Telephone partitioned itself into a regulated part that sold basic network services to downstream retail

[35] The number of LATAs grew to 193 as non-RBOC areas were added.
[36] Rochester Telephone Co. (1993).

carriers, and a competitive part that competed with these carriers free of rate regulation. Several companies began selling local exchange services using the regulated network, including AT&T, Time Warner Cable, Teleport Communications (before its acquisition by AT&T) and Citizens Telecom[37].

More recently, several state commissions and legislatures are considering measures to divest wholesale network services of incumbent local exchange companies (ILECs) from their retail service operations. In a leading case, the Pennsylvania Public Utility Commission has imposed 'functional separation' between the wholesale and retail divisions of state's Bell company, Verizon-Pennsylvania[38]. Other states are currently considering wholesale-retail separations along these same lines.

Clearly, pressure was mounting on administrative and legislative institutions to reform local exchange regulation. The successes registered in federal deregulation of several other network industries – airlines, natural gas transmission, trucking and rail service – set the standard for the telecommunications industry. The successes of individual local competitors demonstrated that facilities-based competition was possible – competition that represented innovative entry and not just cream skimming[39].

Potential models for reform of local exchange regulation were drawn from many corners of the industry. State-level experimentation with telecom regulation offered a range of alternatives, including some highly innovative and radical policies such as deregulation-cum-price caps in Nebraska and Vermont's 'social compact.' On other occasions, the U.S. imported regulatory models from abroad. The best example here is the price cap mechanism applied to British Telecom (BT) in the U.K. After applying price caps to AT&T in 1989, the FCC extended its use to the largest local telephone companies for selected interstate services. Soon afterwards, many state commissions and legislatures adopted some form of price cap regulation to intrastate services. The evolution of this policy now added features such as revenue sharing. During the 1990s, the majority of the states adopted incentive regulation in various forms[40].

[37] As of end of 2000, 11 facilities-based carriers and 3 resellers accounted for 27 percent of the local exchange lines in the Rochester area. See New York Public Service Commission (2001).

[38] "Re: Structural Separation of Bell Atlantic-Pennsylvania Retail and Wholesale Operations," Pennsylvania Public Utility Commission, Docket No. M-00001353, March 22, 2001.

[39] At least in one instance, a backlash against competition was mounted when AT&T sponsored a 1975 bill to roll back competition (the Consumer Communications Reform Act). The bill was withdrawn 1½ years later after intense opposition from new entrants (e.g., MCI, Datran) and large users.

[40] Interestingly, there has been a significant incidence across the states of reversion to traditional RB-ROR regulation. In many cases, reform measures expired and, while a permanent incentive plan was often put in place, on occasion the state returned to traditional RB-ROR regulation. In some cases the reform plan was prematurely terminated. Frequently, the reason given was dissatisfaction with the outcome of incentive regulation experiment, often because of observed poor customer service. See Abel and Clements (1998) and the chapter by Sappington in this *Handbook*.

These reforms were aimed at moving rate levels and structures closer to cost; they were not designed to affect directly the level of competition in local service markets. A series of FCC initiatives took steps toward generating more local competition by reducing entry barriers or otherwise facilitating entry. First, in 1980 the FCC issued its non-dominant carrier order that released qualifying providers from the burden of traditional rate filing and certification of facilities deployment. Second, formation of rules to give competitive long distance companies 'equal access' to local networks, and their implementation after AT&T divestiture, gave ILECs experience in inter-carrier relations. Third, the FCC adopted rules for provision of "comparably efficient interconnection" in its 1986 Computer Inquiry III decision. More groundwork for unbundling the public switched network would be laid when the Commission approved the BOCs' 'open network architecture' proposal in 1990. Finally, responding in part to petitions submitted by competitive access providers, the FCC ordered ILECs to provide facilities-based local competitors 'expanded interconnection' for dedicated and switched services in 1991 and 1992, respectively[41].

At the state level, we saw earlier how progressive states broke new ground in creating institutions to accommodate facilities-based local competition. In New York and Illinois, the commissions and the carriers developed arrangements to permit competitive carriers to interconnect with Bell companies' networks. In the case of Rochester, the New York commission brokered a scheme that provided for wholesale provision of basic local network services as well as resale of local retail services to competitors, and now the Pennsylvania commission has separated the incumbent carrier along these same lines.

Experimentation with various policies aimed at creating local network competition had been incremental and sporadic. However, each experiment added to a national debate which was headed in the direction of a significant, widespread reform of the telecommunications sector. The momentum culminated in passage of the "Telecommunications Act of 1996" (TA96). Fundamentally, the 1996 Telecommunications Act embraced competition and deregulation as the best means to achieve efficiency in local telecommunications markets and to speed widespread deployment of advanced technologies and services. It explicitly rejected regulation as an obstacle to these objectives. The Preamble clearly and succinctly articulates this goal:

> "[The Act will] provide for a pro-competitive, deregulatory national policy framework designed to accelerate rapidly private sector deployment of advanced telecommunications and information technologies and services to all Americans by opening all telecommunications markets to competition..."

[41] FCC Docket 91-141, Interconnection with Local Telephone Company Facilities, September 23, 1991.

TA96 created several new breeds of local network competitors, and facilitated expanded competition in many ways, that we will discuss below in more detail.

2.3. The experience abroad

In some respects, development of the local telephone industry abroad followed much the same early pattern witnessed in the U.S.: a Bell-like monopoly prevailed during which time the first telephone networks were built, first the local exchanges and later the long distance networks. The departure came at the end of the patent monopoly period. In almost every country outside the U.S., the industry was nationalised or absorbed as a government function, typically the post office and sometimes also national banks. These so-called 'PTTs' (for Post, Telephone & Telegraph) were fully integrated into provision of long distance and international services. In most cases, however, they did not integrate into equipment manufacturing as did AT&T[42].

For decades the state-owned telecommunications carriers operated without any threat of competition; independent regulation was irrelevant. The privatisation movement changed all that, and that movement kicked off when the Thatcher government in the U.K. decided in 1982 to sell off British Telecom. Nationalised in 1911, BT was privatised in 1984[43]. Up until 1991 only Mercury Communications was allowed to compete with BT in local and long haul markets. When competition was allowed in local service, it was the nascent cable industry that offered the most direct competition by building their networks to provide voice telephony as well as entertainment video. As of September 2000, U.K. cable firms had signed up over 5½ million business and residential fixed lines providing local telephone service, or 15.8 percent of the nation-wide total[44]. Access provided by fixed wireless technology has met with much less success in the U.K. Up to recently, BT had been banned from both cable and fixed wireless businesses with the goal of promoting alternative access networks.

To regulate rates charged by the privatised BT, the British government devised a system of 'price caps' and charged the newly created regulator, the Office of Telecommunications (OFTEL), with overseeing its implementation. It deliberately rejected U.S.-style rate-of-return regulation as administratively cumbersome and detrimental to incentives. Price caps were applied initially in the U.K. to long haul services. When they crossed the Atlantic in 1989, they were applied by the FCC to AT&T's long distance service first, and then soon afterwards to certain interstate services provided by large local exchange carriers. Some form of price

[42] Nevertheless, the PTTs usually completely controlled nation-wide sale and leasing of customer premise equipment and so held monopsony power over purchase of that equipment.
[43] Although it was not until 1997 that the British government sold off its remaining interest in the company.
[44] OFTEL (2001).

caps, generally referred to as 'incentive regulation,' spread across the country, and each state personalised their implementation by adding certain special features.

On the continent, a variety of alternative competition plans were being implemented, but these were occurring at a much slower pace, usually dictated by the European Commission. That was the case in Germany. The first steps toward competitive telecommunications markets came with liberalisation of CPE in 1988 and the awarding of two GSM licenses the following year, one of which went to the state-owned PTT, Deutsche Telecom (DT). It was not until 1995 that DT was incorporated, however, and then partially privatised in 1996, and it was not until 1997 that an independent regulatory body was created[45].

On the opposite side of the world, a radical restructuring experiment had long been underway. Implementing the Telecommunications Act of 1987, the New Zealand government carved out telecommunications operations from its postal service and banking department, and called it 'Telecom New Zealand' (TNZ). A couple of years later TNZ was privatised and sold to two U.S. RBOCs, Ameritech and Bell Atlantic. Similar to BT, TNZ remained vertically integrated in local, long distance, international, wireless and other services, but not in cable services. While TNZ's privatisation was not extraordinary, the government proceeded to open wide every telecommunications market in the country. Licenses to provide service of any kind were freely issued. No communications-specific regulator was created. Instead, New Zealand's 'light handed regulation' of telecommunications encouraged private negotiations between carriers to establish interconnection arrangements and facility sharing. No pricing methodology or institutions were prescribed to govern these negotiations, only that the outcome did not have dominant carriers (in this case TNZ) violating competition law.

Clear Communications – a *de novo* entrant into long distance service partly owned by MCI initially, and then wholly owned by BT – was the first to build local and long distance facilities and request interconnection with TNZ. The protracted negotiations and ensuing litigation constituted the world's first test of unregulated markets in the mature telecommunications industry. Clear charged that, in demanding rates to terminate traffic on its network, TNZ had used its dominant position to exclude competitors from the long distance market – a violation of New Zealand's 1986 Commerce Act[46]. TNZ replied that its rates were not anti-competitive because they were consistent with the Efficient Component Pricing Rule (ECPR) which had the property of inviting new competitors if and only if they were more productively efficient than the incumbent. The case eventually was appealed to the British Privy Council which concluded that TNZ pricing was not unlawful three years after the initial

[45] Ruhle (1999).

[46] Specifically, Section 36 which prohibited use of a dominant position to deter or exclude competitors or hinder competition.

complaint was filed[47]. Such legal disputes have remained a fixture of New Zealand's telecommunications industry ever since. Litigation arose at each major incursion into TNZ's markets, as Clear moved into local business services, as BellSouth New Zealand entered with its nation-wide GSM wireless network and as Saturn Communication began delivering telephony over their cable system in Wellington. All this may change if pending legislation is passed that creates a telecommunications commissioner in the Commerce Commission[48]. If passed into law, this law would represent a significant reversal in New Zealand's commitment to light-handed regulation.

A radical opening of local telecommunications markets had long before commenced in another part of the southern-hemisphere in a country not much larger in size or population than New Zealand. Chile had pioneered local competition even before U.K. or N.Z. had undertaken privatisation and deregulation[49]. The state-owned local telephone monopoly, Compania de Teléfonos de Chile (CTC), lost its exclusive monopoly in 1979 and two competitors, CMET and CTM, were issued licenses to enter its markets two years later. Chile's 1982 General Law on Telecommunications threw wide open the doors to competition by liberalising licensing of competitors, mandating interconnection, and decontrolling rates.

Akin to the New Zealand experience, competition in Chile was inseparable from litigation. Nearly every attempt at entry was met with private suits, and on occasion, Supreme Court challenges to the authority of Subtel, the regulator overseeing the interconnection arrangements. Unlike the case of the former state-owned incumbents in the U.K. and New Zealand, CTC dominated local and Entel dominated long distance services after their privatisation. In 1994, the two companies were allowed to enter each other's markets, provided they did so with structurally separate subsidiaries and did not exceed certain market share ceilings. The consequence of all these actions was significant competition in all telecommunications market, including local services. As of the end of 2000, at least seven competitors achieved an 18 percent share of local lines in the country[50]. This competition has been facilities-based, with sizeable overbuilds in the Santiago area, and despite the fact that unbundling and resale provisions were only recently put in place.

[47] Telecom Corp. Of New Zealand, Ltd. v. Clear Communications, Ltd. Privy Council, 1994. Note that as a consequence of New Zealand's 1840 entry into the British Commonwealth, it submitted to the British legal system.

[48] Telecommunications Act 2001. Certain 'designated services' can be regulated if the parties to a sale cannot agree on the terms. In making a rate determination for "designated services," the Telecommunications Commissioner selects among the principles of TSLRIC, bill and keep, or a combination of the two. Initial prices for designated services are to be set by benchmarking against rates in comparable countries that employ forward-looking cost methodology or a form of bill and keep. Final prices are invoked only if some party appeals.

[49] For a history of Chilean telecommunications, and a detailed record of the competitive era, see Melo (1998).

[50] Telefónica-CTC, December 2000.

Restructuring and deregulating the world's telecom sectors is a work in progress. There is much that remains to be done before competition is a reality in the many local exchange markets. Nevertheless, the progress has been breathtaking. As one indication, in 1989, 26 of the 29 OECD countries had fixed network monopolies whereas the remainder were duopolies; by 2000, there were no monopolies in this group and 24 now had 3 or more local wireline carriers[51]. The progress in wireless was equally dramatic: none of the monopoly structures in 23 of the OECD countries survived through 1998[52].

3. Economic conditions of local network competition

3.1. Defining local services and markets

Analysis of this industry should begin with a definition of the economic market. For this we use a modified version of the market definition articulated in the Federal Trade Commission-Department of Justice Merger Guidelines: the local network market consists of the smallest collection of communication services and geographic areas that include traditional local exchange services such that a small but significant and non-transitory increase in price (SSNIP) above the *competitive level* will be profitable for a hypothetical monopolist. To make this definition operational, we might begin with all retail switched services, both access and usage, supplied over the wireline network in a metropolitan area, and then consider a 10 percent increase in rates for all those services for a one-year period. If users readily substitute away from such 'plain old telephone service,' or POTS, then the services they switch to should be included in the market definition.

Wireless mobile service is an example of a potentially good substitute for POTS provided its price is not exorbitant. In fact, the wireless alternative raises an important point: the SSNIP profitability test should be performed at competitive rate levels, and not current rate levels. POTS rates are likely below costs due to political desires to cross-subsidise local service. In comparison, for much of its history, cellular service has been priced high relative to its costs and relative to basic wireline service, making it less attractive as a substitute for POTS. For users, wireless may be unattractive relative to POTS, not because it is not functional (most would agree that wireless has greater functionality), but because it is too pricey.

On the supply side, mobile service providers could offer local exchange services in competition with wireline networks, but not without building new capacity in wireless local loops. Cable television networks may be in a better position to roll out service to the general population in response to a price increase. Cable telephony is being offered by most of the largest multiple system operators (MSOs) in

[51] OECD (1999 Table 1.1) and OECD (2001 Table 2.1).
[52] OECD (2001 Table 1.1), op.cit.

the U.S. Whether it will scale up to the broader population in a reasonably short period of time remains to be seen, and so whether cable telephony offers a substitute for POTS on the wireline network remains an open question.

Increasingly, telephone companies and their competitors such as cable operators offer customers a package of services. Bundling blurs market boundaries because users may choose not to switch to avoid forgoing perceived benefits of one-stop shopping. Digital convergence – meaning the combination of different voice, data and video services on the same physical medium – allows carriers to expand their range of service offerings at relatively small incremental cost compared to building a stand-alone network. As a result, an increase in basic local rates may not cause a customer to switch to wireless mobile service because local service includes high-speed Internet access. On this flip side, many entrants – especially cable companies – view the possibility of offering customers multiple services over their network as a means to pry customers away from the incumbent, especially when incumbents cannot respond in kind due to line of business restrictions.

Demand-based and supply-based delineation of local markets can be correlated. As a given set of individuals become more dispersed over a given geographic area, the cost of supplying them will most likely increase. At the same time, their demand for communication may also increase since the alternative of face-to-face contact becomes more costly. In part, this was one of the reasons that, quite surprisingly, sparsely-populated rural areas of the U.S. had some of the highest rates of telephone penetration in the early days of the industry.

Determining the geographic boundaries is an essential exercise to focus analysis on local telecommunications markets and differentiate from long distance and other services. Strictly speaking, users desire connections between specific originating and terminating points, so that these unique pairs define a unique service which does not have close substitutes. A slight change in terminating station will invariably result in no value to the user as in the case of dialling a wrong number. Adopting this definition of a service, however, makes the overall number of services astronomical[53].

Individuals do not wish to communicate with every other individual on the planet, but they do place calls throughout the country and possibly the world, and they do value the option of being able to call any number sometime in the future. Because people tend to have more numerous social and commercial relationships with individuals and businesses that are close by, a vast majority of calls are made within the local geographic area. This is especially the case of local businesses that have a physical local presence, e.g., banks, retailers, utilities, schools, and governments. Of outgoing residential calls in the U.S.,

[53] As an example, if there are n users, then there are $n\ (n-1)/2$ bi-directional pairs, and so with a billion terminating lines worldwide we would have 500 quadrillion unique products.

84.4 percent are local, 2.8 percent are local toll and 12.8 percent are long distance[54]. For this reason, in the aggregate, we may take geographically local markets as service provided within a contiguous population centres, such as a metropolitan statistical area. Census Bureau and Rand McNally definitions of population centres better approximate such markets than the LATAs drawn by the MFJ court.

Strictly speaking, the term local varies from one individual to the next: it will be an area in which their home (and possibly their place of work) is centrally located. With wireless mobile service, the position of the individual is no longer fixed, in which case the centrally located position may well be along a highway between home and work and shopping areas. When delineating the market, we will consider a change in price of local service throughout a region and the aggregate response of individuals living in that region.

Cost of supply of local services is very dependent on the geographic extent of the market. This is certainly true for traditional wireline networks where the length of the loop directly determined the cost of providing a user with access. Other technologies make distance less relevant. The cost of providing fixed wireless access, for instance, does not vary by the distance between the transmitting tower and the customer location up to the point where the signal attenuates. Also, with a ring architecture, service becomes unrelated to distance since traffic between two neighbours on the ring may travel its entire length.

With the proliferation of service options, substitution patterns among offerings become more complex. Connections may differ in their content (voice, video, image/fax, data), mobility (stationary, nomadic, high speed mobile), and bandwidth. As an example, wireless mobile services (e.g., PCS) have recently closed the gap with wireline service by adding custom calling features (call waiting, call forwarding, ANI), paging, voice mail, and now email and web pages. As a result we can expect greater substitution away to wireless should the price of wireline increase.

One final way to delineate local markets is by the customer type. The crudest distinction is between business, residential and carrier. In some cases there is little distinction as when either businesses or households wish to subscribe to mobile wireless service. But in other instances, one customer has no demand for a service that the other values highly: e.g., households have no demand for PBX trunks, and neither households nor businesses purchase unbundled local loops like competitive carriers. Even that distinction has blurred as the ranks of self employed individuals working from home increase. Each category also has important distinctions due to both demand differences and the cost of serving them. For instance, there can be huge cost differences in serving urban and rural households, and scale economies to serving large businesses.

[54] FCC, 1999 Statistics of Communications Common Carriers, August 2000. Note that no account is taken of type of call (voice, fax), or its distance, duration, and time of day.

Vertical services are available that could displace some local *usage* in response to a price increase, though they all may employ local wireline *access*. One such service is voice messaging. Whether a carrier service or user supplied (i.e., an answering machine), voice messaging may substitute for calling – unless parties engage in telephone tag. Facsimile transmission is another service that could substitute for local usage; it is also likely to stimulate local usage as users substitute away from regular and express mail services and email. Email itself may provide an alternative for sending certain messages locally. It has been estimated that 55 percent of emails displace voice calls, though these are not necessarily local calls[55]. Finally, 'instant messaging' (IM) could eventually offer an Internet version of dial tone, as it eliminates the latency associated with e-mail and yet captures a community of interest[56]. Surprisingly, the 'dual system' which predominated in the era of local phone competition, persists today with IM as AOL deliberately blocked compatibility with other IM networks[57].

How do we measure local exchange competition? Whatever yardstick is chosen, end users should be the judges. If they get lower prices, better quality, greater variety and innovation, then market conditions are competitive even when the structure may remain quite concentrated. We might divide the different answers to this question into measures of inputs and outcomes. Among the inputs are structural and behavioural conditions. Outcome measures can be divided by whether they track the success of entrants or weigh the competitive impact on incumbents.

We lack subscriber data sufficiently detailed to compute traditional structural measures of concentration for each local network market such as the HHI. CLECs currently offer fixed telephony service in 56 percent of zip codes which amounts to about 88 percent of the U.S. population[58]. Nevertheless, active local lines provided by CLECs number about 16.4 million as of December 2000, or a mere 8.5 percent of the U.S. total[59].

In crafting the TA96, Congress deliberately refused to measure competition in terms of market share, the number of providers, or any other quantitative yardstick. Alternatively, we might look for behavioural conditions that are conducive to competition. A good example of this approach is the 14-point checklist for local competition found in Section 271 of TA96. Among other items, the list requires incumbents to supply competitors with physical interconnection and

[55] See Anderson et al. (1995) citing an Electronic Messaging and Micro Systems journal study in Huber (1987).

[56] Nick Wingfield, Changing Chat: Will Instant Messaging be the Dial Tone of the Future? Wall Street Journal, Sept. 18, 2000.

[57] In the Matter of Applications for Consent to the Transfer of Control of Licenses and Section 214 Authorisations by Time Warner Inc. and America Online, Inc., Transferors, to AOL Time Warner Inc., Transferee, FCC Cable Services Docket No. 00–30, Memorandum Opinion And Order, January 11, 2001.

[58] FCC, Local Telephone Competition: Status as of December 31, 2000, May 2001.

[59] Ibid.

non-discriminatory access to rights of way, poles and conduits. These conditions do not ensure competition will occur; they are believed to increase the likelihood of entry when they are met[60].

It has become typical to measure competition in terms of the acquisition of market share by entrants. This can be measured by more traditional indicators of competitor success such as sales and profit. In the specific case of local competition we record the number of phone numbers ported to competitors, collocation agreements and number of wire centres covered, unbundled elements supplied (loops, ports, trunks, switching) or resold lines. Alternatives to such output measures of competitor success are amounts of investment in fibre miles, customer lines, and switches installed. By one measure, CLECs had 16 percent of local fibre miles as of the end of 1998, but only 1.8 percent of the customer lines and 2.4 percent of local revenues[61]. In addition to the fact that CLECs must build infrastructure in advance of signing up customers, this mismatch attests to how incumbents can benefit from customer inertia.

A weaker structural test would assess evidence of potential competition. This would be the number of potential rivals in adjacent markets such as cable and wireless service providers, electric power utilities and incumbent local exchange carriers who are located in nearby territories. It would be important, however, to assess the degree to which these companies are committed to competing in this market, recognising that to sink investment in a physical network that extends throughout the local area (as with a cable TV franchisee) is quite different from merely reselling an incumbent's retail services.

Once faced with effective competition, we would expect incumbent monopolists to lose market share to entrants and to see their profits fall, as prices and sales are driven down. A by-product of this competitive pressure could easily be to spur the incumbent to cut its costs, improve its service quality and accelerate new product innovations. Tomlinson (1995) and Woroch (1995) offer some evidence that ILEC service reliability improved after entry of CAPs.

3.2. Demand for local network services

An analysis of market demand for local services contributes to an understanding of this industry in several ways. In the past, demand analysis aided regulatory commissions in determining the proper level and structure of local service rates. Price elasticities were used to make tariff adjustments and, more recently, cross-price elasticities guided the rebalancing of rates. Access elasticities are useful when evaluating initiatives to achieve universal service. As we enter a deregulated, competitive era, demand analysis is needed for other purposes. Cross elasticities

[60] The Act allows ILECs seeking to enter interexchange service to follow "Path B" in which competitors have not materialised despite the fact that the checklist is satisfied.

[61] Federal Communications Commission, Local Competition: August 1999 (1999, Chart 2.1 and Table 3.1).

now help in the definition of local service markets, a first step toward distinguishing services that are competitive from those that are non-competitive. When competition results in multiple viable suppliers, it is now important to measure *firm* demand functions (as opposed to market demand functions) to gauge, among other objectives, the extent of market power of current carriers and the effectiveness of competition.

Before launching into an evaluation of various demand studies, a few distinctions affecting local services are needed up front. First, it is important to distinguish between access to, and usage of, local services. Carriers compete differently in the two dimensions, as when wireless mobile matches wireline service by offering big buckets of minutes that begin to approximate flat rate tariffs. Second, we can differentiate basic services from 'enhanced' or 'vertical' services. Basic service refers to the collection of services that makes up POTS: dial tone, a phone number, local switched service, long distance access, white and yellow directories, directory, repair and emergency services, and billing. Enhanced services include custom calling features such as call waiting, call forwarding, number identification, plus voice mail and Internet access. Finally, users may need to buy necessary equipment (telephone handsets, inside wire, PCs, facsimile and answering machines, set-top boxes, software) in addition to the service itself.

Since most demand analyses are based on market data, it is necessary to understand how local services are priced. In the U.S., most users subscribe to basic local service for a fixed monthly fee that includes unlimited calling within the local area[62]. The amount of the flat rate varies depending on the size of the local calling area selected by the subscriber. Typically, ILECs are required by state commissions to set the same rates throughout their serving territory. An alternative to flat rate service is measured service, which has the subscriber paying a smaller monthly fixed charge, and an additional amount based on usage (measured by calls or minutes or message units), usually with some calling allowance. Measured service is an option for residential users throughout the U.S., though in New York City and Chicago it is mandatory. Business users are usually charged on a usage basis for their outgoing calls, and pay significantly higher monthly subscription fees compared to residential users[63].

To get an impression of access and usage of local services, the average U.S. household spent $398 in 1999 on basic local service, or about half of the $830 spent on all telecommunication services[64]. This figure translates into 1.12 percent

[62] An alternative to the subscription system is prepaid service, a billing method that has become popular for mobile wireless and long distance services but not for local service. As an example, OFTEL (2000a) reports that 3 in 5 mobile customers in a U.K. survey use prepaid service.

[63] This may or may not be price discrimination since, although the services are identical technically, it costs more to serve businesses due to their higher calling volume that typically occurs during peak hours. Volume discounts built into multi-line plans and Centrex and PBX trunks may compensate large businesses for high mark-ups on their basic service.

[64] Federal Communications Commission, Trends in Telephone Service (2000).

of average American household expenditures on all goods and services in that year[65]. At the end of 1999, 193.9 million local loops were in operation in the U.S., of which 127.8 million were residential. In total, 99.1 million American households had telephone service by July 2000, making for a national penetration rate of 94.4 percent[66]. The average U.S. household with telephone service had 1.21 fixed lines[67]. During 1998, the average American local loop had 60 minutes of use per day, where 45 minutes, or 75 percent, was local calling[68]. As of the end of 2000, there were over 109 million mobile wireless subscribers, so that, on average, roughly 40 percent of the U.S. population had mobile wireless service[69].

Given the opportunity, researchers would prefer to examine access and usage decisions by conducting a controlled experiment in which businesses and households faced various service and price options. Users would choose among two of more service providers representing various technological alternatives, and each would offer different calling plans, contract terms, and service packages. To get directly at the issue of local competition, we would like to track purchase behaviour over the time when the industry transitions from monopoly to more competitive structure. Such an ideal experiment has not been conducted, and will likely never occur. Instead there is much to learn from patterns of consumer behaviour under traditional monopoly structures, but also from early experience with local competition and with wireline-wireless competition.

In a classic econometric study of local access, Perl (1983) estimated the demand for subscription to the PSTN. Using data from the Census Bureau's 1980 Public Use Sample, Perl estimated a discrete choice model of demand for access based on rates (at the wire centre level) and demographic characteristics. He found probit coefficients on monthly subscription rates in the range of -0.0175 to -0.0492[70]. Income effects were relatively more sensitive compared to the price effects, having a probit coefficient of $+0.1296$, while the installation charge effect was quite inelastic at -0.0034.

The study of access demand can answer questions about telephone penetration and the most effective means to promote universal service. Quite apart from any goal of distributional equity, one reason to promote widespread access to the local telephone network is to take advantage of 'network externalities.' These occur when each subscription confers a benefit on all existing subscribers because they can now call, and be called by, the new subscriber. This so-called "network externality" is increasing in the number of users connected (as well as their

[65] Ibid.
[66] Ibid.
[67] Federal Communications Commission, Trends in Telephone Service (2000), op.cit. Note that this figure ignores residential telephony delivered by other media such as a part of cable modem service.
[68] Federal Communications Commission Trends (2000) op.cit.
[69] CTIA (2001) and U.S. Bureau of the Census, Monthly Estimates of the United States Population: April 1, 1980 to July 1, 1999, with Short-Term Projections to November 1, 2000.
[70] These are logit coefficients; not usual price elasticities.

intensity of use). Certainly in the early days of the telephone industry users considered who might they be able to call when subscribing to phone service. Similarly, users who lived in cities with non-interconnected dual systems were very concerned about relative sizes of the competing local networks. More recent examples of network externalities in the communications industry include the adoption of facsimile machines and email services.

In his study, Perl (1983) found that demand for residential access was increasing in the density of phone subscription in a household's local calling area, confirming the presence of a network externality. The effect was small, however, as might be expected given the high U.S. telephone penetration rates during the sample period. Furthermore, unlike the earlier competitive experience, all networks were interconnected, further realising the available network externalities. Nevertheless, new services always appear on the horizon, some of which may substitute for local service and they may not be completely interconnected, as in the case of instant messaging.

Whether a household has phone service is not a terribly interesting question given the high penetration rates in the U.S. Instead, the question is what 'portfolio' of access lines a household will choose. A typical household may likely have two wireline phones (often with a second line reserved for a fax machine, Internet access and/or teenagers) plus a cellular phone and cable TV service. Each of these access media is a potential or an actual substitute for the others to the extent that they all can be used for voice and data communication, and in some case, video as well. Facilities-based competition, at least initially, is likely to supplement, and not entirely replace, current access lines to the household as usage is diverted from one medium to another.

Demand for access media other than the PSTN represent potential sources of competition such as fixed and mobile wireless access and cable TV service. When evaluating the competitiveness of these alternative access media, we must recognise that they could exhibit some network effects that will have the tendency to slow their penetration rates early until they approach critical mass. Relatedly, a new service (e.g., instant messaging over web-enabled wireless phones) will tend to obey the typical S-shaped diffusion curve as users learn about the technologies, incorporate them into their daily lives, and make necessary complementary investments that replace existing equipment and services. In both cases, the observed migration from incumbent network and reductions in usage are at least partially independent of the relative pricing of the services.

Turning to local usage, the scarcity of measured service in the U.S. makes for limited data to estimate price and income effects. An exception was a two-year field experiment conducted by GTE in 1977 which sought to analyse household response to mandatory measured service in three small towns in central Illinois. Residential subscribers paid varying rates for each call and for each minute of a call. Monthly data were gathered on both the number of calls and the total minutes of use by each household. Using these data, Park, Wetzel and Mitchell

(1983) estimated the per-call price elasticity of number of calls to be -0.076 and the per-minute price elasticity of usage as -0.055.

Responsiveness to usage-sensitive local rates have a profound impact on the take up rate of new services and, hence, revenue models of new communications ventures. Arguably, the U.S. witnessed rapid Internet adoption through dial-up modem connections, in part, as a result of the prevalence of flat-rated local service. In time, ISPs also adopted flat rated pricing. An immediate consequence is that the typical call to an Internet service runs about five times as long as the average voice call. In contrast, Europe (and other regions) have lagged in Internet adoption and usage, in part, because users must dialup over measured local service. Flat rated or free ISPs provide European Internet users with some relief.

Local usage is dependent on who is paying for the call. Both parties likely benefit from a call, so that if just one party pays, it confers a 'call externality' on the other[71]. Typically, wireline service adopts a 'calling party pays' (CPP) system that charges the party who initiates the call. Toll free numbers and collect calls reverse the charges to the called party. An important exception to CPP is wireless service in the U.S. where the mobile user pays airtime charges for both incoming and outgoing calls. It has been argued that this system is one reason for the slower penetration rate of wireless service in the U.S. – despite its earlier introduction – especially relative to Scandinavian countries and Japan[72]. It is also likely to retard the spread of wireless Internet usage relative to CPP countries.

To date, econometric demand studies have focused, out of necessity, on local access and usage purchased from a regulated monopoly provider. As a consequence, these studies are silent on a number of important issues raised by local competition. For instance, the typical demand model in the Perl tradition lacks prices of substitutes whether they were the new access methods, such as cable telephony, or more traditional alternatives like paging or payphones or cellular mobile. It is too early to point to empirical models of consumer substitution among local providers but there is a body of evidence that has emerged that looks at competition among providers of short haul (interstate, intraLATA) toll. Taylor (1999) finds an intraLATA demand cross elasticity of $+0.23$[73], indicating that residential consumers are willing to shift usage away to some extent from their local exchange provider to a competitive intraLATA supplier. In study of fixed-mobile substitution, Ahmad, Ward and Woroch (2000) find household willingness to shift local toll usage from wireline to mobile wireless service. Neither of these local toll studies examined substitution among access alternatives. A promising research area is an understanding of consumer local access and usage response when presented with competitive alternatives.

[71] A call externality would occur with data services when a dialup connection would benefit both the client and the server even though just the user or the Internet content provider may pay for it. See Taylor on "Customer Demand Analysis" in this *Handbook*.

[72] See also the chapter by Hausman in this *Handbook*.

[73] Taylor (1999, p. 13).

This exercise will be complicated by the fact that, as a consequence of competition, carriers offer users a plethora of calling plans and service packages. Enabled by digital convergence, and pushed by competitive urgencies, carriers have begun to offer business and residential users bundles of communications services at attractive rates relative to purchasing the individual services separately. In a competitive environment, it will be crucial to understand the ability of entrants to break into local service markets by offering a bundle of services, and of incumbents to retain customers in response to competitive threats. In an early contribution to understanding this strategy, Kridel and Taylor (1993) estimate consumer response to the bundling of two custom calling features. More complicated than consumer response to a package discount, however, is the extent of consumer interest in purchasing multiple services from an integrated supplier, rather than various specialised providers. If there were evidence that consumers valued 'one-stop shopping,' then we would have some basis for this popular strategy that has led to huge cross-media mergers and expenditures on network retrofits.

This brings us to the issue of consumer willingness to switch providers when competitive alternatives come available. Earlier we discussed various sources of consumer inertia such as cost of switching phone numbers. Policies will eliminate some of these costs as when number portability is completely implemented. Many other sources of inertia remain, however, and it is important to determine the extent of this inertia to, among other objectives, evaluate the feasibility of competition and its likely pace. In a series of reports, OFTEL in the U.K. has conducted several consumer surveys to better understand the extent of switching and its determinants, within and between fixed and mobile alternatives[74]. Predictably, these studies find that the prospect of reduced charges was the main enticement for consumers to switch providers or technologies, and that satisfaction with current supplier is the principal reason for loyalty. Another area of demand analysis opened up by local competition is the study 'dual subscription' to multiple providers[75].

3.3. Cost of local service and technical change

The technology of local service provision, and the resulting costs of production, are instrumental in determining whether competition is feasible and desirable in local service markets, and what form it will take. It is common to refer to the amalgam of various parts that deliver the traditional voice and data services as the PSTN.

At a basic level, local service provision, like so many communication services, is delivered by a network that is composed of links and nodes over which traffic of

[74] Oftel (2000a,b,c).

[75] Oftel (2000d pp. 4–5) finds in one survey that 7 *percent* of households with a fixed line subscribe to two fixed-line local providers, more often than not BT and a cable operator.

various kinds travels. The nodes of the network are individuals with their phones, faxes, personal computers with modems, and answering machines, and similarly, businesses with their PBXs, LANs, and email and web servers. The public network is full of nodes as well; these are made up of central office, remote and tandem switches, and main distribution frames and cross connects. The links that connect the various nodes are the copper loops, coaxial and fibre cables and radio connections. The carriers supply the inter-office trunks, feeder and distribution plant, and inside wire. Information that is electronically encoded as analogue or digital signals travels over the network. This information may be voice, data, facsimile, or video; it may be one-way or interactive, or broadcast; it may be switched (either circuit or packet) or unswitched.

Historically, the PSTN has been viewed as monolithic in the sense that a single provider builds, operates and maintains the network from end to end. Certainly this is the model that AT&T successfully promoted under Theodore Vail's direction and that the PTTs pursued outside the U.S. The PSTN was nevertheless designed to be modular in the sense that it could be decomposed into parts that readily inter-operate, provided the parts adhered to Bell interface specifications. The network was not open, however, in the sense that a third party could unilaterally attach equipment to the network or run traffic over it. AT&T turned to outside vendors for some pieces of equipment but only when they met AT&T's engineering standards. By and large over time, AT&T ferociously guarded access to its network as when it banned placing plastic dust covers on directories arguing that this was necessary to ensure network integrity! Under these conditions, competition will occur only if a competitor builds an alternative network and poaches the incumbents' customers. An alternative, if not complementary approach, is to open up the existing network and allow competitors to purchase its services as inputs to their offerings.

3.3.1. Scale and density economies

Communications networks experience strong scale economies, in part, a consequence of the relatively large fixed cost associated with establishing links. Wireline links require one-time expenses of acquiring rights of way, burying conduit, erecting poles and stringing cable. Wireless links have even stronger scale economies: once spectrum is acquired and a transmitter tower is in place, variable cost is negligible for reaching customers within the range of wave propagation. Beyond the costs of establishing a given number of connections, we might expect production to obey constant returns to scale as the network is replicated to cover a wider territory[76].

[76] 'Lumpiness' of network investment poses a countervailing force to constant returns to scale. Many components of modern networks, from fibre sheaths to digital switches to transmitter towers, are available in indivisible, minimum sizes. As a result, networks are built to a capacity that exceeds current demands and excess capacity persists.

Scale economies are indicated when average cost exceeds marginal cost. Before checking this relation, a preliminary question is: What is the basic unit of output to measure the cost of local services? It could be an additional minute of talk time or an additional call; it could mean a second line to a household or adding a household to the network. Whether empirically there are scale economies will be sensitive to the choice of units.

Often it is claimed that there are no variable costs to delivery of local services. This is wrong for several reasons. First, even while the overwhelming majority of costs are durable investments, the plant and equipment have very sharp capacity limits, and as those limits are approached, shadow cost of capacity expansion rises[77]. Second, the variable costs of running a local network entail customer acquisition and care. Typically, customer acquisition costs in competitive cellular telephone markets range from $300 to $500 per subscriber. The importance of these costs grow as customer churn increases and the average tenure of a customer falls, as will likely occur as local services become more competitive.

Within a given territory, however, the investment cost per wireline will tend to decrease with population density. As density increases, the average loop length falls as nodes on the network become more closely packed. Of course, this requires that switching and terminal expenses do not outstrip the savings in transmission investment.

Switching and transmission investment substitute for one another[78]. Consider the simplest 3-node network. Complete service can be achieved by building a link between all three possible pairs of users. This 'mesh' architecture requires investment in three links and no switching. Alternatively, a switch can be installed at one of the nodes, allowing the network to eliminate the link not connected to the switch and re-route traffic between that pair of nodes indirectly through the switch. The load on the remaining two links, of course, will be heavier by the amount of the indirect traffic. Alternate routing of this sort economises on transmission and is the basis for the circuit switched network[79]. Airlines realised these economies when they adopted their current 'hub and spoke' route structure.

With large shifts in the relative costs of transmission and switching, the architecture of telephone networks adjusts to economise on construction expense. In the 1880s when switching was manual and cumbersome, wiring was so heavily used in large cities that they blocked out sunlight in the densest areas[80]. In the 1980s when electronic switching became increasingly affordable, and while fibre optic transmission was still costly, telephone companies began to migrate toward a double-star network topology – at least when wiring newer neighbourhoods.

[77] See Woroch (1987) for an expression for this cost. Mitchell (1990) provides engineering estimates of this shadow cost.

[78] See the chapter by Sharkey in this *Handbook*

[79] See Sharkey (1982).

[80] The urban blight that resulted led to bans on above-ground wiring in New York City as mentioned above.

By locating remote switches and digital loop carriers between central offices and customers, and by replacing the feeder plant with high capacity trunks, the carriers substitute switching for transmission.

More generally consider design of a network to connect n nodes. A complete network with a separate link connecting each pair of nodes would require $n(n-1)/2$ links in total, but no switching. If, instead, a single switch is installed, then only n links are needed, one loop connecting each node to the switch. The question then becomes whether the savings of $n(n-1)/2 - n = n(n-3)/2$ links exceeds the additional cost of a switch (plus the additional cost of necessary transmission capacity on the n links to handle the greater traffic).

The trade-off between transmission and switching is dependent on more than the number of nodes and the relative cost of switches and cable. It also depends on the dispersion of users. To see this, consider a 4-node network in which users are located at the corners of a rectangle with height L and width $1/L$. By design, there are four users per square unit. With L near 1 (*i.e.*, a square), it makes sense to locate a single switch at the centre. As L gets large (*i.e.*, the rectangle becomes elongated with two pairs of users separated from one another), and as long as there is no cost of capacity nor any capacity limits on the links, it becomes economical to install a second switch, locating the two switches at the far ends of the serving area[81]. Here we see how a disperse population justifies multiple switching centres connected by high capacity trunk lines, as is typical of the PSTN. It also illustrates how population dispersion can present profit opportunities for entrants who can efficiently serve individual switching centres provided inter-exchange trunking is available.

3.3.2. The natural monopoly question

The main reason for our interest in scale economies of local services is to help determine whether the industry is a natural monopoly, and so evaluate claims that competition will harm overall social welfare[82]. The cost-based definition says natural monopoly prevails if one firm can provide all amounts of service at a lower cost than could two or more firms. Formally, production cost must be subadditive at each level of output within the relevant range[83]. This is a strong,

[81] If a network has a single switch, it is optimal to place it in the centre of the rectangle, in which case the cost of building the network will be: $S + 4F + 4c(L^2 + 1/L^2)^{\frac{1}{2}}/2$ where S is the cost of a switch, F is the fixed cost and c is the unit cost per mile of installing a line, and $(L^2 + 1/L^2)^{\frac{1}{2}}/2$ is the length of each local loop. When there are two switches, they are located between the two close users at the short edges of the rectangle. Then the network cost is: $2S + 5F + 4c(1/2L) + cL$ where a fifth trunk line connects the two switches. Two-switch architecture is less costly than a one-switch architecture when the transmission cost saving exceeds the cost of the second switch and the fixed cost of the interoffice trunk: $2(L^2 + 1/L^2)^{\frac{1}{2}} - (2/L + L) > (S + F)/c$. A necessary condition for this to occur is that $L > 2/3^{1/2} \approx 1.155$ so the population dispersion does not have to be terribly great to justify a second switch.

[82] Faulhaber (1975).

[83] Baumol (1977).

global test for monopoly to be the cost-efficient industry configuration. It is also static in that this condition is checked at a point in time, and assumes cost is not dependent across time.

It is felt that the scale economies that derive from large fixed costs of building different parts of the telephone network, especially local loops, feeder plant and interoffice trunks, ensure the local exchange will have costs that are (globally) cost subadditive. Evans and Heckman (1983) were among the first to empirically test for natural monopoly in telephone services. They analysed pre-divestiture AT&T using a two-product translog cost function, where the products are local and long distance services. Testing for sub-additivity of the cost function using 1958–1977 data, they failed to reject (local) super-additivity and they could not accept (global) sub-additivity. Extending Evans and Heckman's data to 1979, Röller (1990) estimated a CES-quadratic cost function – again for local and toll services – and accepted global cost subadditivity. Using accounting data submitted by the large local operating companies to the FCC over the pre-divestiture period 1977–83, Shin and Ying (1992) estimated a translog cost function. They found evidence of weak scale economies at the central office level, and concluded costs were not subadditive[84].

It must be kept in mind that these and other econometric studies employ accounting cost data collected from highly regulated telephone companies. As usual, accounting measures depart from economic cost concepts. More important, these regulated companies are likely driven off their efficient production frontier by regulatory constraints[85]. These carriers are also protected from actual and potential rivalry that might otherwise drive them to be more efficient.

3.3.3. Scope economies

Another important economy in local service provision is the savings from delivering multiple services from the local network. A distinguishing characteristic of the local network is the fact that it provides access to all kinds of service besides making local connections. Consequently, it is more efficient to share one local network across these services than building multiple networks. The industry learned quickly the high cost of this structure during the dual system era.

Scope economies prevail when it is cheaper to provide two (or more) services by a single entity than by two (or more) firms each specialising in a single service[86]. At the source of the scope economy is a shared facility which, in the case of the local network, are various types of real estate (rights of way, radio spectrum, central office floor space), infrastructure (ducts, conduits, poles), transmission facilities (trunks, loops, towers) and switching equipment (including routers and servers).

[84] See also Shin and Ward (1993) and the Fuss and Waverman chapter in this *Handbook*.
[85] When telephone providers are state owned, no obvious objective motivates their supply behaviour.
[86] Panzar and Willig (1981).

Credible tests for scope economies in local network services are few. Several of the econometric studies of scale economies mentioned above test for the presence of scope economies. In each case the fitted cost function is simply projected out of sample to predict costs of a stand alone local service provider. Gabel and Kennet (1994) avoid this problem by employing an engineering optimisation model to estimate costs of providing local services. They find strong scope economies among switched services, as well as between switched and non-switched services.

One advantage of so-called 'digital convergence' is the ability to realise scope economies by combining several different transmissions on the same medium (when they had previously been carried on separate facilities)[87]. Currently voice and data share parts of the local network (loop and trunks, but not switches or servers). The overlap between voice and video on the cable network is another example as voice simply occupies low-frequency spectrum on the coaxial cable. Fixed and mobile wireless services have much less overlap, with wireless using the landline transmission facilities for traffic backhaul and long distance[88].

Whereas the presence of scope economies argues for a single firm to produce a range of services, the benefits of specialisation can work in the opposite direction. Specialists can avoid the overhead needed to deliver a wider range of services and yet enjoy large-scale economies. It is a theoretical possibility and a practical reality that specialists find entry into narrow product markets profitable. The technical condition that holds when a service is vulnerable to specialised entry states that the entrant's stand alone cost of producing any level of the service up to market demand is less than the incremental cost to the incumbent of adding that service to its product line.

Common carriers are especially vulnerable to 'cream skimming' by entrants who offer services with the largest price-cost margins. Their supplier-of-last-resort obligation compels them to serve all customers under the prevailing tariff and to provide a full range of basic services throughout their serving area. On the other hand, users have identified benefits of buying from one supplier in the form of reduced transaction costs in billing, customer service, and technical support. So there may be economies of specialisation in production but benefits from one-stop shopping on the buyer side.

One last form of scope economy worth mentioning takes the form of cost savings when a firm integrates across two successive stages of production[89]. Such 'vertical economies' likely arise in the local services industry for provision of network carrier services (unbundled elements, bulk transport) and retail services (basic services, private line), and in turn, basic local service with certain vertical services (customer calling features, voice mail). As with horizontally related

[87] Katz and Woroch (1997) describe several other interpretations of the concept of digital convergence.

[88] Although a wireline network may terminate incoming wireless calls, and vice versa, provided the two networks are interconnected..

[89] See Spulber (1991).

services, the presence of shared resources across stages is a prime source of vertical economies. Sharing infrastructure – both physical and software – can reduce the overall cost of a package of wholesale and retail services.

It appears, however, that the savings to integrating upstream into network construction and equipment manufacturing is not great. It is in these situations that a specialist in a particular stage of the supply chain can make profitable entry. It offers a sound explanation for why local telephone carriers often outsource such activities as inside wire maintenance and billing and collection. There is also an important strategic reason why vertically integrated service providers have been divesting upstream operations as competition materialises downstream. AT&T divested Western Electric, now Lucent, mainly because its customers saw a conflict of interest buying equipment from a company with whom they competed in services.

An important property of telecommunications cost, especially infrastructure investments, is its 'sunkness.' Once completed, much equipment and facilities have little economic value in uses other than for what they are intended. Fibre cable buried beneath the street or pulled through underground conduit is literally and economically sunk.

Sunk investment has important strategic implications for local network competition depending on its size and who makes the investment. When incumbents invest in sunk facilities, they have incentives to cut price down to avoidable costs. Technologies that do not involve heavy sunk costs can be much more responsive to uncertain, changing conditions. For instance, wireless technology tends to be less sunk than wireline technology[90]. To the extent that transmitters can be moved or the system can be retrofitted for another wireless standard, wireless technology is cheaper to use. In contrast, cable TV companies have learned that a network optimised for one-way delivery of multi-channel video entertainment is not easily reconfigured to deliver two-way voice and data services.

3.3.4. Technical change

Technology of local telecommunications, like any communications industry, is improving rapidly and unpredictably. What is certain, however, is that costs are falling and capabilities are expanding. Advances in microelectronics have lowered switching costs as price and performance of essential digital signal processing (DSP) and application specific processor (ASP) chips steadily improve. The relentless march of improvements in optical transmission has increased the throughput of a fibre while lowering the cost of manufacture and installation of fibre. As the performance-adjusted cost of switching and transmission fall together, how the trade-off discussed above will balance out remains to be seen.

[90] An exception would be the investment in certain kinds of spectrum licenses. If those rights were not transferable to other carriers and/or not portable to other locations, then the expenditures become more sunk, and accordingly they would fetch lower bids.

More important for our purposes, however, are the impacts that changes in scale economies have on local competition. Roughly, technological change has reduced overall costs of providing local services, but with an increase in fixed costs relative to variable costs. On balance, the effect on competition is ambiguous. An entrant that adopts new technology, thereby facilitating entry can achieve lower costs; but the relatively higher fixed costs raise MES, reducing the number of viable firms. To complicate matters, new technologies often result in improved or enhanced service stimulating demand and likely supporting more firms.

Digital switching and transmission have profoundly reduced costs and improved service of local communication networks. Digitalisation supports advanced signalling networks and switching features that expand the services offered to customers. Digitalisation also elevated the role of software in the local network. A typical digital switch is run by programs having more than a million lines of code enabling an expanding array of capabilities. Compression algorithms pack increasing amounts of information on an optical fibre (e.g., dense wave division multiplexing) or on a radio channel (e.g., 3G wireless standard), greatly relaxing capacity limits.

Compared to early times, the local network is not a permanent, static platform for service delivery. Local carriers can reconfigure switching and signalling software to launch new calling features unanticipated when the switch was first installed. In the case of business services, this control may be placed in the hands of the user. Carriers can also trouble shoot and maintain their networks electronically from a central, regional operations centre, reducing the necessary 'truck rolls.' As a result, reliability and security of wireline and wireless networks have greatly improved[91].

The impact of innovations in network architectures on local competition is more difficult to predict. An example would be the deployment of fibre rings in urban areas rather than the traditional 'star network' with loops emanating from central office switches. Certainly this new architecture was a good fit for CAPs' strategy to enter by gathering high volume traffic in a densely populated business area. These networks economised on switching – relatively recently installing carrier-grade switches – taking advantage of the low cost and high capacity of fibre.

More recently carriers have been moving toward Internet Protocol (IP) architectures patterned off the way the Internet is configured. Whereas the traditional PSTN centralised network intelligence (switching and signalling), the Internet places intelligence at the edges of the network. The Internet is said to be made up of dumb pipes and smart terminals, compared to the PSTN where the terminals are dumb and intelligence resides in the network.

Packet-switched networks use of alternate routing is pervasive but with different implications for switching and transmission. In that case information including voice and other traffic are 'packed' at the source, and each packet is

[91] Though centralised control, without adequate redundancy, can also raise reliability risks.

individually routed to its intended destination where it is reassembled in its original form. Compared to a circuit-switched network where a dedicated circuit is established between the source and destination, packet switching reduces the need for capacity along a route between the two nodes. The dynamic routing of packets allows the traffic to reach its destination when congestion or failure makes some routes unavailable. So far, synchronisation problems, high latency and poor security have limited the use of packet network to carry voice traffic.

Econometric studies using historical data cannot address questions related to the current and near term technologies of the local exchange, when technologies change so fast. Alternatively, engineering cost studies offer the promise of a look at the near-term future. By using current or projected prices for inputs and by adopting the best available technology and optimal architectures, these models aim to better predict costs that incumbents and entrants face at present and will face in the immediate future.

As with econometric studies that use data from monopoly industries, engineering cost models may be based on costs that reflect an imperfectly competitive industry. Pressures of competition will affect pricing of equipment and other inputs, and also result in different equilibrium network architectures. In fact, engineering cost studies have been most prominently built and used for regulatory matters. For instance, the FCC developed the 'Hybrid Cost Proxy Model' to generate cost estimates for unbundled network elements (UNEs) which, in turn, would be used to set prices for these services[92].

Current demand and cost analyses fall far short of informing the most urgent questions surrounding local network competition. Out of necessity, econometric modelling of these phenomena has had to rely on data drawn from incumbent monopolists and state-owned enterprises whose rates and service offerings were highly controlled. In the new era of local competition, what is needed are estimates of the costs of entrant supply as well as those for an incumbent who faces competition. This may involve a greenfield construction of a network, or it may rely on unbundled network services to varying degrees. The latter represent altogether new wholesale services that must be added to the incumbent's product line. In addition, the entrant will likely offer a different array of services than the incumbent. It is the reality, however that, as of yet, competition has been too limited and too brief to provide an adequate dataset to test these kinds of propositions.

4. The structure and regulation of the local network industry

4.1. Structure of the U.S. local exchange industry

Supply of local telephone service has always been highly concentrated in a given geographic region, even during the early era of dual system competition. Today in

[92] Sharkey (1999).

the U.S., and increasingly elsewhere in the world, the emerging structure is one of a dominant incumbent facing an array of facilities-based and service-based competitors, including *de novo* start-ups as well as established firms entering from related network industries. Prominent among this latter group are long distance carriers, cable television systems and electric power utilities. Cable and electric power have facilities that overlap considerably with local exchange networks. The serving area of the large cable multiple system operators (MSOs), and major long distance carriers, also extend beyond regional presence of even the largest ILEC.

Concentration in local services can be measured against several alternatives, and depending on what segments are examined, the extent of competition can vary considerably. This is due to the fact that competition has taken hold in certain customer segments, service offerings and urban areas. The penetration of competition continues to grow in these markets, and has spread to other markets over time.

As of end of 1999, the four RBOCs together owned 88 percent of U.S. end-user lines with Verizon and SBC accounting for nearly two-thirds between them[93]. The RBOCs collected 94 percent of ILEC end user revenues[94]. While there are hundreds of small and mid size local exchange companies, they tend to have even larger shares in their markets than the RBOCs since their markets often can not support multiple carriers. At the close of 2000, CLECs provided 16.4 million lines or what is 8.5 percent of the 193.8 million of the nation's fixed lines[95]. CLECs owned 7.75 million of those lines, acquiring the remaining two thirds from ILECs as UNEs or resold loops. In 1999, CLECs as a group has sales of $6.5 billion, and had a cumulative average growth rate exceeding 87 percent over the proceeding eight-year period[96].

The extent of local competition is greater if one were to treat mobile and fixed wireless services as competitors. Measured in terms of revenues, cellular and PCS carriers in the U.S. generated over $52.5 billion in 2000,[97] a figure that represents a third of the combined local wireline and wireless communications sales[98]. Now since many ILECs also own wireless carriers in their wireline serving territory, this last figure must be reduced to better reflect the relative size of wireless sector as a competitor to fixed line service.

The picture of concentration is much more varied if one looks at individual local markets. At one extreme, as has been typical, are the communications-

[93] Federal Communications Commission, Trends in Telephone Service (December 2000, Table 8.3).
[94] Federal Communications Commission, Statistics of Communications Common Carriers (2000, Table 2.9).
[95] Federal Communications Commission, Local Telephone Competition (2001).
[96] Federal Communications Commission, Local Telephone Competition at the New Millennium, (2000, Table 7).
[97] CTIA (2001).
[98] Calculated using Federal Communications Commission, Statistics of Communications Common Carriers (2000 Table 2.9) op.cit.

intensive markets of the largest metropolitan areas. By the end of 2000, CLECs reported 2.95 million lines in New York state, accounting for 20.9 percent of the lines in that state[99]. Fully 93 percent of zip codes in New York state are served by at least one CLEC, and as many as 32 percent of them are served by 7 or more CLECs. This compares to a nation-wide figure of 56 percent of zip codes being served by at least one CLEC[100].

Relative success of competitors also varies across customer types. At the broadest level, CLECs have been much more successful in winning business customers away from the ILECs. Only 41 percent of CLECs' lines are sold to residential and small business customers compared to 79 percent of ILECs' lines.

Competitive carrier share continues to grow at a rapid pace, starting from insignificant levels 10 years ago. During the year 2000, the number of CLEC resold lines and leased UNE loops nearly doubled while the number of ILEC end-user lines actually fell by 2 percent[101]. Nevertheless, the four RBOCs and the other ILECs have well over 90 percent of combined business and residential switched access revenues[102].

The facilities-based competition that has materialised to date has employed traditional copper loop technology. By the end of 2000, of the 16.4 million competitive lines in service, only 1.1 million were delivered by coaxial cable and another 451,000 by fixed wireless[103]. Together, these two alternative technologies represent just 10 percent of CLEC lines and less than 1 percent of all end-user lines in the U.S.

At the same time that entrants into local network markets multiply and grow, the concentration among established carriers has greatly increased. A series of mergers transformed the largest eight U.S. ILECs into just four. In the two most prominent consolidations, SBC acquired, in order, Southern New England telephone, Pacific Telesis, and Ameritech, and Bell Atlantic acquired Nynex and then GTE to form Verizon. By and large, these mergers were among contiguous regions, laying the foundation for providing long distance service over a wide region. The effect has been to increase concentration among local exchange incumbents at the national level and at the same time to increase scale and potential of threats to those markets. Both the SBC and Verizon mergers were approved with requirements that the companies meet a timetable for out-of-region entry into local exchange markets, with stiff financial penalties for missed deadlines.

Over the past decade, the long distance industry has invested significant sums of money to acquire facilities-based carriers who provide, or could provide, local

[99] New York Public Service Commission (2001).
[100] Federal Communications Commission, Local Telephone Competition (2001, Table 12).
[101] Federal Communications Commission, Local Telephone Competition (2001, Table 4).
[102] Derived by combining interstate switched access and state access revenues reported in Federal Communications Commission Statistics of Communications Common Carriers (2000, Table 2.9).
[103] Federal Communications Commission, Local Competition, (2001, Table 5).

services. AT&T made the largest expenditures on local service infrastructure. The company purchased one of the country's two largest CAPs, Teleport Communications Group and its largest cable company, TCI, plus a portion of MediaOne, a cable giant. MCI purchased the other large CAP Metropolitan Fibre Systems (MFS), in 1996 for $14.1 Billion, before buying two smaller ones, Brooks Fibre Properties and Intermedia Communications. The third largest long distance carrier, Sprint, purchased United Telephone, the independent local telephone company.

All three major interexchange carriers also took a large stake in wireless service, both fixed and mobile[104]. The cable industry took steps to prepare for its attack on the local markets. MSOs replaced their trunk network with optical fibre, and bought and sold or swapped properties to create contiguous clusters of franchises in urban areas. Attempts to join cable and local exchange companies were broadly unsuccessful and rarely used as a means to enter out-of-region local exchange markets.

Established carriers, both incumbents and entrants in local services markets, are groping toward the most effective scale and scope. AT&T has come full circle in this regard: after being divested in 1984, the company entered nearly every facet of the local wireline and wireless market through acquisitions. Its expressed motive was to offer customers the full array of services with the convenience of one-stop shopping. Recently, AT&T has chosen to divest itself into four companies, each pursuing separate business lines.

Entrants into these markets have gone through several evolutions, first focusing on dedicated access services, and then switched services, and now Internet services such as web hosting. And whether incumbent or entrant, wireline or wireless, all carriers recognise the growth potential of data services, and the limitations of voice telephony, and to a lesser extent video services. The ongoing explosion in growth of data traffic could overshadow other threats to the PSTN. As usage shifts away from circuit-switched networks onto packet-switched facilities, revenue growth from Internet access could ensure the viability of cable, fixed wireless and satellite networks that could bundle in basic voice telephony at a negligible incremental cost.

To summarise, local network markets have historically been highly concentrated. One hundred years ago, in the first competitive era, residential and business customers had a real choice among carriers, although they sacrificed benefits from network externalities since competing networks typically were not interconnected. Local service customers, especially residential, have had much less choice in the recent episode of local competition, but they have realised the full benefits of interconnected networks. Five years after complete opening of these

[104] AT&T acquired the largest cellular company at that time, McCaw Cellular, and then later Vanguard Cellular. The other two major inter-exchange carriers went for fixed wireless acquisitions. MCI bought the fixed wireless firm CAI Wireless and Sprint bought People's Choice TV, American Telecasting, Wireless Holdings, and Videotron Bay Area.

markets, competitors have achieved roughly 6.4 percent of local revenue and lines[105]. This progress has been considered inordinately slow by many, but it must be put in perspective. By 1981, 10 years after it first began service, MCI had achieved a mere 1.05 percent share of domestic toll revenues[106].

The reality is that implementation of competition involves a detailed, costly restructuring of a mature industry which consumes a considerable amount of time. Incumbent carriers did not have any significant experience with arranging inter-carrier sales of wholesale network services to competitors. Indeed, such inter-carrier transactions did not occur even in those services which were arguably competitive prior to deregulation. As such, the interconnection, unbundling and resale that underpins much of the current level of competition are unfamiliar and unnatural acts for the ILECs.

4.2. Regulation of local network competition

No portrayal of the structure of the local network industry and its patterns of competition would be complete without discussing the regulatory policies imposed on these markets. Until recently, regulation was almost exclusively an issue for the U.S. telephone industry since elsewhere local networks were state owned. World-wide privatisation of PTTs has spawned new regulatory agencies and institutions, and as a by-product, we have a richly varied experiment of alternative policies toward local competition.

Regulation of local network competition takes on many different forms. Some are direct in their impact on competition by facilitating or impeding entry, while others are more indirect by restricting how incumbents (and entrants) can compete in local services markets, and consequently their incentive to enter in the first place. Regulators face at least four distinct challenges, the combination of which is somewhat unique to local network competition.

First, they must decide whether structural competition is beneficial or whether, perhaps because of natural monopoly conditions, entry is not viable and potentially wasteful. Policy makers cannot know the efficient amount of competition without full knowledge of the cost conditions facing established firms and potential entrants. Difficulties in making this assessment are magnified when entrants (or incumbents) deploy entirely new technologies that lack any track record on costs.

Second, as with many network industries, competition in local services may require some time to grow to a size necessary to realise scale economies in both supply and demand. In the meantime, an entrant may require assistance or protection if it is to achieve viability. In comparison, incumbents have invested in local network facilities long before competition materialised. Economically sunk

[105] Local Telephone Competition, op. cit.

[106] MCI's 1981 revenues stood at $413 million while total toll sales were $39.18 billion. See Trends in Telephone Service, FCC, Feb. 1995 at Table 30.

and ubiquitous, these networks can selectively and rapidly respond to entrants' forays with in any specific area.

Third, efficiency demands that services be provided over shared facilities, as happens, for instance, when local and long distance voice and data traffic travel over the same local loop serving a residential customer. The presence of shared facilities raises the issue of whether relative prices are structured to subsidise one service over another. The proper test is whether a service generates unit revenue that falls between its average incremental cost and its average stand-alone cost. A regulator seeking to subsidise some service or customer class, perhaps to pursue universal service, invites or deters entry in selected markets. One might question to what extent entry by CAPs was merely a response to cross subsidies built into switched and special access rates.

Fourth, and related to the previous point, sharing of facilities by different service providers can have a strong efficiency appeal. Even if two local providers do not overbuild one another, efficiency demands that they terminate each others' traffic, for otherwise users would be denied full benefits of the network effects. Pricing the use of these facilities – whether for the terminations of traffic or the leasing of facilities – must balance the incentives of competitors to enter local markets against the incentives for incumbents to maintain and replace existing facilities and to build new ones. In particular, should incumbent carriers be permitted to recover some or all of the historical costs of the facilities they are required to offer competitors? We do not address this difficult, crucial question in this chapter, or other issues related to the efficient level of interconnection prices[107].

Agencies empowered with regulating local services have over time arrived at a division of labour. States have jurisdiction over local telephony and intrastate toll services, while the FCC regulates interstate toll and related services. Conflicts between state and federal regulators arise in large part because local facilities are shared by the different kinds of traffic. As one example, the FCC has decided that if more than 10 per cent of the traffic on a local facility was interstate, then the entire facility was treated as falling in the FCC's rate regulation domain. Using this rule, many CAPs were able to claim FCC jurisdiction and get out from under more restrictive state regulation. Increasingly, local governments play an important role in controlling local entry through their control of urban rights of way, conduit and ducts, and often impose franchise taxes on competitive local carriers[108].

[107] See the chapters by Armstrong and by Noam in this *Handbook*.

[108] The connectedness of various parts of the local network can cause deregulation to have unintended consequences. Often opening up one part of the product line to competition will create pressure on adjacent services. For instance deregulation of long distance created incentive to cut local access charges by spawning bypass operations and encouraging IXCs to integrate upstream into provision of local access. Deregulation of CPE created new means for users and competitors to bypass ILECs: the PBX substituted for Centrex, microwave links bypassed special access circuits, answering machines substituted for centralised voice messaging, and more recently, personal computers with modems facilitated "voice over Internet protocol" (VoIP).

Cable television services also are principally the responsibility of the municipal authorities who awarded their franchises. The federal government did not become involved in the early days of the cable industry, with a few key exceptions. In 1972, the FCC ordered that new cable systems must be provisioned for upstream channel (anticipating cable telephony services)[109]. The agency imposed its ban on cross ownership of cable and telephone companies in 1984, with the primary goal of keeping telephone companies out of video delivery. Congress also took the initiative to intervene in constraining basic cable rates. With passage of TA96, these rates were de-controlled in 1999. Even today, the regulation of cable is segregated from other communications services such as wireline telephony and wireless services in most regulatory agencies (e.g., Wireline Competition, Wireless Telecommunications and Media Bureaus at the FCC).

Supply of local telephone service has been treated as a *de jure* if not a *de facto* monopoly franchise by state authorities. Perhaps the lingering memory of dual local systems compelled the agencies to effectively block entry into switched local services to residential customers. These regulatory barriers are necessary to maintain low local rates deemed necessary to promote universal service, among other objectives. A complex system that combines rate of return rate making with fully distributed cost pricing drives a wedge between local service prices and their costs. It is generally accepted that, to varying degrees, intrastate and interstate toll rates have been a significant source of subsidies that finance the shortfall[110]. Within the local exchange, there is evidence that business access and usage rates subsidise residential rates,[111] dedicated access charges subsidise switched services, and custom local services, operator and directory services and yellow page directories all contribute to subsidies that support lower local rates. Pressure develops for entry where regulated margins are greatest, and because of economies of density, this usually occurs in the most populated urban areas.

Another important aspect of local rate making is the asymmetric treatment of incumbents and entrants. At least for interstate access services, entrants enjoyed non-dominant status which often meant they merely filed their tariffs for public inspection. Restriction on incumbents' rate setting can severely limit their ability to respond to competitive threats, and artificially inflate profitability of entry. The pricing flexibility that comes with many "incentive regulation" schemes re-balances this asymmetry to some extent.

[109] Cable Television Report and Order, 36 FCC 2nd 143 (1972) required that all future cable systems be designed to be two-way capable by setting aside a 25 MHZ band for "return traffic."

[110] In the opposite direction, urban local rates clearly subsidise rural local rates if only because many states require geographic averaging even while costs of service are much higher in small towns and rural areas.

[111] Palmer (1992) finds that business basic service rates subsidised local service in Massachusetts.

TA96 takes significant steps to encourage entry by facilities-based local service providers with the following measures:

1. Mandating interconnection of networks at technically feasible points, including physical and virtual collocation of network equipment,
2. Furnishing other resources and services that aid in entry including phone numbers, dialling parity, rights of way, ducts, poles, databases, directories,
3. Reciprocal compensation for transmission and termination of calls across networks,
4. Eliminating legal barriers to entry into local telephony by: (a) relaxing the FCC's cable-telco cross-ownership ban; (b) pre-empting state and local regulations from creating entry barriers; and (c) exempting electric power utilities from restrictions by the Public Utility Holding Act.

The Act also opened entry routes into the local exchange by service-based providers by requiring:

1. Unbundling of network elements and supply at rates near economic cost (TELRIC),
2. Resale of retail services at wholesale prices equal to retail rates less avoided cost.

Entrants may also use any combination of these services to enter local exchange markets, along with their own facilities.

At least initially, the Act has succeeded in stimulating entry of both kinds. Between the first quarter of 1996 and the end of 1999, the number of CLECs holding numbering codes increased from 16 to 275, with the numbers ported going from zero to 4½ million[112]. Over this same period, the combined number of unbundled local loops and resold lines has gone from zero to 8.3 million[113].

To further encourage the RBOCs to open their regions to competition by implementing these measures, the Act extends the "carrot" of permission to enter in-state long distance service. An RBOC may get this LOB restriction lifted for a state provided they satisfy each item on a 14-point checklist as well as gaining approval by the state commission and the FCC. After several attempts, the first RBOC to succeed in getting FCC approval was Bell Atlantic in New York state in 2000. The long distance market has been opened to RBOCs in several more states since that time.

Whether the provisions of the Act will ultimately succeed in creating effective, sustainable competition for ILECs will not be known for years. After all,

[112] FCC, Trends in Telephone Service, March 2000, Chart 9.4, and Local Competition: August 1999, report by Industry Analysis Division of the FCC's Common Carrier Bureau, August 1999.
[113] FCC, Local Telephone Competition: Status as of June 30, 2000,Table 1.

facilities-based long distance carriers and resellers numbered in the hundreds for many years after the opening of this market, even while AT&T never held less than half the market 30 years after the landmark MCI decision.

TA96 has left its mark on local communications markets abroad as well. It provides a model that other regulatory authorities imitate, and learn from. Recently, the U.K.'s OFTEL imposed local loop unbundling on BT[114] and the European Commission separately embarked on specification of a similar policy[115]. These initiatives come after years of promoting infrastructure competition over service-based alternatives. It raises the question as to which form of competition delivers the highest social returns. The U.S. does not offer a clean natural experiment since local markets were opened to both forms of competition at the same time. In fact, a convincing interpretation of the Act's promotion of unbundling and resale of local services is as an attempt to provide entrants a stepping stone to full facilities-based entry. Arguably, platform competition that could emerge in local service markets – with the alternative platforms being the public switched network, the hybrid fibre-coaxial cable network, the mobile and fixed wireless networks, and the electric power grid – is the only means by which widespread, durable competition will survive.

5. Strategic modelling of local network Competition

5.1. Causes and consequences of local network competition

5.1.1. Meaning of competition

Before discussing the causes and consequences of local network competition, it is important to determine what is meant by competition. An increase in the number of firms serving a local service market, net of any exit, is usually interpreted as an increase in competition[116]. Of course, the larger the new entrants – measured by their initial scale or the depth of their financial pockets – the greater the competition they contribute to the industry.

Empirically, the life span of a typical entrant tends to be fleeting. Geroski (1992) finds that, as a group, entrants into a wide range of industries take considerable time to accumulate a rather small market share. So far, this generalisation has been supported by the experience in local service markets. With the passage of time, and with accumulated investment, an entrant

[114] OFTEL (1999).

[115] European Commission, "Proposal for a Regulation of the European Parliament and of the Council on Unbundled Access to the Local Loop," 12 July 2000, Brussels.

[116] While exit reduces this number we should note that assets of bankrupt providers that can be sold to subsequent entrants, and used by them, can speed the build out required for successful entry.

graduates to become an incumbent; the speed and extent with which this occurs is important in these markets because each new wave of local service competitors may rely on the facilities of earlier generations to gain entry.

Apart from structural measures, competitiveness of these markets will vary with price and non-price rivalry among local service providers. Price rivalry becomes more aggressive, for instance, following the removal of regulatory restrictions on rates charged by local providers. Rivalry may also intensify when users treat products and services offered by these firms as closer substitutes. This occurred when cellular telephony closed the quality and reliability gap with wireline alternatives, and the two technologies were drawn into head to head competition[117].

The level of competition might also derive from disparities in carriers' costs. When, through investment, a local provider succeeds in lowering its costs, it becomes more competitive vis-à-vis its rivals. Sometimes merely reducing the employment rolls brings about the reduction in costs. In an ironic turn, by laying off employees, incumbents seeking to be more competitive have, in turn, added to the pool of potential entrepreneurs that has fed telecom start-ups[118].

5.1.2 Causes for local competition

As evident from the above discussion, two key sources for local network competition are the presence of enabling technology and accommodating regulatory policy. In certain markets, the existence of unserved demands for local service is another inducement for competitive entry.

Restructuring of established carriers may cause local network competition. Often this may occur under legal pressure as when AT&T agreed to divest to avoid further antitrust action. That divestiture created seven local telephone companies where there had been just one, and also a powerful potential entrant in the form of AT&T's long distance division. Other restructurings are voluntary; once again, AT&T provides an example. The company has begun to divest its various lines of business including its broadband cable telephony and wireless divisions.

Technology played a central role in local telephone competition. Expiration of initial Bell's patent set off a land rush among prospective licensees vying for the most lucrative urban markets. After an initial chapter of competition with Bell, Western Union ceded local telephony (and local telegraphy) services to the Bell

[117] In another example, CAPs initially targeted markets that were under-served by ILECs using innovative technologies. In time they expanded their offerings to more traditional markets and standard technologies, and competition between CAPs and ILECs intensified generating benefits for users in the form of lower price and improved quality.

[118] Many of the largest new ventures into telecommunications have been founded and managed by alumni of AT&T, including Qwest, Teligent and Global Crossing.

interests, preferring to focus on what it viewed to be more lucrative long distance markets[119]. Once the Bell patent expired, a second land rush broke out as independent telephone companies descended on markets served by Bell operating companies, as well as smaller markets and rural areas outside the Bell ambit. Scale diseconomies of early switching systems was one of the most significant limitation on the size of the region that local companies could profitably serve[120].

Easily a more significant factor in creating competition in a network industry like local telephone service is the possibility that entrants can interconnect their network with incumbents and otherwise gain network services that allow them to reach scales that are needed to achieve viability. These rarely occur voluntarily but rather are the product of regulatory and legislative mandate.

Competition, driven by desire to control local phone traffic, as well as the need to control costs, led to the invention of the automatic switch. Strowger, an undertaker, who suspected that the local switchboard operator was routing calls of potential clients to her husband (also an undertaker), allegedly invented one of the first automatic telephone switches. By automating a labour-intensive activity, incumbents who adopted this technology became more competitive, and the potential for over builders evaporated. In comparison, innovation aided new entrants when CAPs built network control centres to monitor and repair their networks from a central location with minimal need for field workers.

These examples show that innovations have indeterminate impacts on competition. In some cases they enhance scale economies and the cost advantage of embedded networks, augmenting the dominance possessed by an incumbent. In other cases they facilitate entry at small scale with low capital outlays and/or a highly differentiated service. Whereas the innovations may lower overall cost, MES could rise or fall, and affect concentration accordingly. A by-product of fast paced innovation is rapid obsolescence. Consequently, the advantage of the current incumbent is limited by the next generation of equipment that replaces it within a short span of time.

Standardisation of technical specifications can not only realise significant scale economies in production of the hardware and software needed to build the local

[119] Friedlander (1995, p. 29). Bell's telephone was capable of providing telegraph services as well as voice telephony. Even while Western Union concluded that local telegraphy was not worth the trouble, Bell's invention represented an early example of infrastructure convergence that brought new and established firms into competition with one another. At that time telegraphy had less severe distance attenuation problems than telephony. But with advances in copper wiring and the invention of the loading coil, long distance telephony would supplant Western Union's principally interurban network of iron and steel wiring which suffered from special limitations of its own. See Fagen (1975, pp. 200–209). In the largest urban areas, the telegraph network had a counterpart to the local exchange in the form of a system of pneumatic tubes that delivered paper messages to telegraph offices. See Standage (1998).

[120] Scale economies were also limited by the fact that early telephone systems were powered by batteries at both the customer premise and the central switchboard, a problem that was solved by the 'common battery switchboard' that centralised the powering of the network.

network, but it can greatly ease the burden on entrants. A communications platform built on publicly available interfaces stimulates competition around the edges of the network as specialised firms are able to deliver products and services that inter-operate with the public local network.

Digital convergence holds the promise of expanded competition in local network markets. Exploiting the digitalisation of switching and transmission, local telephone companies can invade cable's video delivery market, cable systems can provide two-way voice services, and ISPs can use packet switched technology to carry voice conversations over the Internet. In the case of cable telephony, the incremental cost reduces to retrofitting the system to handle two-way digital signals and the installation of switching equipment. To be balanced, given that these networks were initially optimised for a specific service (e.g., one-way delivery of analogue multi-channel video), some compromise must be accepted relative to greenfield construction of a specialised network. At an operational level, its technicians must gain expertise necessary to accommodate a wider range of services on the networks they maintain.

Regulation can induce entry by creating cross subsidies in local rate structures. The lucrative markets beckon new competitors who pray the high margins will last long enough for them to recoup their investment, or to move into other markets. Such pricing could induce inefficient firms to enter the market, but if there are plenty of potential entrants, the more efficient ones should win out.

Asymmetric regulation can also facilitate entry competition with established carriers. This occurs, for example, when rules applied to incumbents are more restrictive than those for younger or smaller firms. Entrants may simply be relieved of the burden of filing and justifying their rates or their investment plans. A wider wedge is driven between incumbents barred from entering certain lines of business and entrants who enjoy much greater freedom, with the belief that, only when protected from competition, will new firms enter and thrive. In particular an incumbent is deprived of bundling an extensive range of services that might otherwise allow it to retain its local service customers.

As discussed in the previous section, TA96 obliges incumbents to supply entrants with network services with the goal of promoting competition. Pricing of these carrier services is crucial to the success of this approach to competition and has been the locus of intense legal and regulatory confrontation. Deregulation of rates and entry is a more pro-active means toward competition. Besides eliminating artificial barriers that block entry by efficient competitors, the TA96 raised maximum foreign ownership percentage with the effect of expanding the pool of competitors.

Relaxing rate regulation can also intensify price competition, as when cost-based rate regulation of incumbents is replaced by incentive schemes. The expanded freedom tends to eliminate any price umbrella that may shield emerging firms from the full effect of competition. Over the longer run, this freedom can have a depressing effect on competition as judged by the ranks of new entrants.

An example is Telecom New Zealand which, while its markets were opened to competition, was given almost complete pricing freedom (as well as minimal obligations to interconnect with competitors) to meet the competitive threats.

Privatisation of PTTs has been the first step toward local competition in many countries outside the U.S. Exposed to the forces of the financial markets, the restructured national carriers have turned in a remarkable record of cost cutting and quality improvement[121]. In other cases, however, local markets were opened to competition prior to privatisation. As signatories comply with the terms of the World Trade Organisation's (WTO) 'Basic Agreement on Telecommunications,' and as European countries adopt the EC's unbundling directive, these markets will witness continued pressure from competitive local companies.

Finally, demand conditions have also played a role in stimulating local competition. Especially outside the major industrial economies, user frustration with poor service quality and long waiting periods for connection has created profit opportunities for entrants able to persuade regulatory authorities they should enter. And even in major markets, certain niche services – typically cutting edge business services – provide openings for competitors, as when CAPs delivered highly reliable, dedicated services to customers who might otherwise have built a private network to obtain the desired service levels. In the data services area, the explosive growth in usage of the Internet and corporate intranets has driven demand for high-speed data access. Widespread diffusion of personal computers, as well as enterprise networks and web servers, fuels the data traffic, as do the growth of electronic commerce and other Internet applications.

5.1.3. Incumbent responses to competition

Incumbent responses to increased competition – though the record is far from complete – reveal a number of patterns. First, when able, incumbents cut rates for their services threatened by competitors. Second, they implement measures to improve their offerings; this might take the form of network modifications that increase capacity or reliability, or it might be more responsive customer service (order taking, billing, provisioning). Third, there is evidence that competitive pressures induce incumbents to cut their costs, usually by paring back their employment roles[122].

Responses by U.S. local telephone companies to threats posed by competitive access providers illustrate several of these points. To begin with, the ILECs cut rates for the dedicated access services targeted by CAPs[123]. ILECs also devoted considerable effort to improving the reliability of their dedicated networks, and to shorten the provisioning time in line with CAP offering. While ILECs were early

[121] Ros (1999).
[122] See Ros (1999) for an empirical analysis at the country level. Gort and Sung (1999) look at long distance productivity using local services as the benchmark for comparison.
[123] Tomlinson (2000).

to deploy fibre ring networks, most of these replaced interoffice transport facilities. They played catch up with the CAPs in terms of rolling out SONET rings in response to CAP competition[124].

One last incumbent response bears mentioning although it is a political reaction to competition. Faced with competition, incumbents have appealed to regulators and solicited the courts to relax restrictions on their responses and to ease their obligations to assist new entrants. Incumbents do not have a monopoly on this political manoeuvre, however. Entrants have gone before regulators as well seeking protection from subsequent entrants once they succeeded in establishing themselves in the market[125].

5.2. Strategic choices of local network entrants

Competitors in local service markets have various options for providing service. Two broad categories of strategies are facilities-based and service-based entry.

5.2.1. Facilities-based entry

At one extreme a firm may provide services over facilities that it owns. Invariably, these facilities are interconnected with other local networks with whom they exchange traffic, as well as with long distance networks – though dual systems of the nineteenth century demonstrated that it is possible for two separate local networks to vie for the same customers. Another more recent illustration is the presence of 'bypassers' that connect business customers to their long distance providers using microwave or fibre optic equipment without ever travelling over local network facilities.

A *de novo* entrant may be a start-up or it might be a firm established in another market that is diversifying into the provision of local services. Fixed wireless provides an example of technology supporting new entry into local access services. The latter may be market extensions (when a cable TV operator gets into the provision of voice telephony). In this way, long distance carriers have entered local markets recently in the apparent hopes of leveraging their existing facilities and telephony expertise.

Alternatively, the foray of an existing firm into an adjacent market constitutes a geographic extension; this occurs when in those rare instances a local exchange carrier attempts to provide service in another ILEC's franchise area. Alternatively, the firm may acquire an established firm or purchase its capital equipment. Certainly buying out an established domestic carrier is a rapid and apparently less costly means for a foreign carrier to enter a domestic market. In the U.S., in an attempt to enter local services markets, first the largest cable

[124] See Woroch (1995) and Tomlinson (1995).

[125] As an example, in 1989, Teleport Communications asked the Massachusetts Public Utilities Commission to deny MFS certification in Boston. See Tomlinson (2000, pp. 189–190).

operators purchased the major CAPs, and then more recently the major long distance companies bought them out[126].

Existing networks, designed for altogether different purposes, can be reconfigured to carry traffic in direct competition with incumbent local networks. Cable telephony offers a prime example. Electric power companies not only use their internal fibre networks for local transmission services, they have also begun to deploy 'powerline' technology that delivers switched voice and data services directly to the home over the electric grid[127]. The incremental cost of adding local access service over these infrastructures may be quite low, and yet their coverage is nearly ubiquitous, and they have in place many existing customer relationships and brand recognition.

5.2.2. Service-based entry

Alternatively, an entrant may rely on the services of the incumbent local network, choosing to own almost no network facilities of its own. Such a service-based entrant not only gains access to essential rights of way and infrastructure, it also may lease wholesale services such as transmission capacity from local electric, gas or water distribution utilities.

The TA96 greatly expanded service-based options to include the purchase UNEs and the resale of retail services of the incumbent, and in some cases, joint occupancy of its facilities as well (e.g., sharing frequencies on the local loop). In implementing the Act, the FCC has defined seven network elements and proposed that the ILEC be obliged to combine them into platforms upon request[128]. If resale is the chosen option, any service the ILEC offers at retail must be available to the CLEC at wholesale rates.

Competitors may choose a mix of different facilities-based and service-based strategies – an option provided them by the TA96. Indeed, one principal motivation for UNEs is to encourage entrants to combine them with facilities that they build or acquire. Different methods may be used in the different regions, where a new carrier can purchase network elements in those markets where it is (and may always be) uneconomical to build a second network. Resale, in particular, may allow an embryonic carrier to quickly achieve a broad footprint as it builds out its network. In these ways UNEs and resale can function as an interim stepping stone to facilities-based entry.

[126] See Woroch (1996) for an evaluation of the relative merits of acquisition strategies for entering the local exchange.

[127] Another surprise comes from below ground where robots are used to string optical fibre through city sewer and water systems. See John Schwartz, "Wiring the City: Humans Won't Do," New York Times, March 8, 2001, p. D1.

[128] The Court of Appeals for the Eighth District upheld the FCC's designation of UNEs. That court also concluded that the FCC exceeded its authority when requiring that the ILEC provide a UNE platform, but the U.S. Supreme Court subsequently overturned this ruling.

Entrants commonly undercut the prices of incumbent carriers, at least when they first appear in the market. This is a necessity when its services are not greatly differentiated from the incumbent's. For this very reason, entrants usually strive to distinguish their offerings from what is currently available. Some variation is achieved by differences in structure of pricing of local services. During earlier episodes of competition, local franchisees were likely to use measured service while entrants quite effectively penetrated the market with flat rate prices[129]. In addition, since it seems that one-time connection charges loom large as a deterrent to switching to a new service, entrants have often held those down, usually amortising them over time or collecting them on other services.

5.2.3. *Evaluating entry strategies*

Both strategies for entering local network markets have their advantages. Facilities-based entry is very costly and, because of the sunkness of the investment, it comes with a high level of risk. On the other hand, entry of this sort poses much more potent competition than the service-based alternatives. When they design, build and own their own facilities, competitors have much greater control of operating costs and definition of services.

Entry by leasing incumbent facilities and reselling incumbent services, in comparison, requires much less upfront outlays, and if those wholesale services are available, can aid the entrant in getting to market more quickly. A service-based entrant, however, has costs that are highly dependent on incumbent pricing of the requisite services and, equally important, the incumbent's design of the features of those services and where and when they are available. In the end, a service-based entrant forfeits considerable control over what it can offer its customers.

In choosing which route to follow, both the existing structure of local service markets as well as its inherent assets enter into a potential entrant's decision. The incremental cost of building a new network is important for the facilities-based alternative, along with the difficulty of integrating it with existing infrastructure[130].

Whether they enter *de novo* or diversify from other markets, and whether they build their own facilities or purchase network services from an incumbent, the more successful entrants into local services follow an evolutionary approach. Typically, they establish a toehold in some services, selling to certain customer segments in some geographic area before venturing into other markets. In the beginning CAPs entered as "carriers' carriers" by hauling long distance traffic among interexchange carriers in the largest urban areas. After initial success, they began to offer dedicated services to large businesses and government agencies. It would take time before CAPs as a group would begin to serve mid-size and

[129] Gabel and Nix (1993).

[130] For an analysis of the different options facing cable operators seeking entry into telephony, see Woroch (1997).

smaller businesses in second and third tier cities, and to provide switched local services to its customers including residential users. Today, CAPs serve a wide range of markets with the fastest growing being in Internet services such as web hosting and caching.

5.3. Strategic models of local network competition

A logical approach to modelling local network competition, given the small number of competitors, is to use game theory methods. This formulation would include at least two types of players, incumbents and entrants. In a simple version, a single incumbent faces a single potential entrant in each period. Strategies include the decision of where to build a network and what capacity it should have, which services to offer and their prices, and whether to interconnect with other networks and how much to charge to terminate traffic. Features of local network competition that we have discussed elsewhere that are less easily incorporated into these models include the presence of network externalities and the sunkness of network investment.

5.3.1. Co-operative game approach

In this approach to network competition, a player is synonymous with a traffic flow that travels along various paths connecting two nodes. The question is whether, for some allocation of the surplus generated by the network, every coalition of players (traffic flows) can be discouraged from breaking away and serving themselves using a standalone network. Usually, for reasons of scale and scope economies, it is assumed that the grand coalition can generate the greatest aggregate surplus for the players taken together. This condition alone is insufficient to prevent a subset of players from defecting. Hence, in the words of co-operative game theory, we seek a 'core' of the network game.

Sharkey (1991; 1995) has made significant contributions to answering this question and identified operations research literatures that address similar questions. In the end the message is not entirely optimistic. Fairly plausible conditions characterising local network environments fail to imply a core. Sharkey finds conditions such that there is nearly a core of the co-operative game. Bittlingmayer (1990) and Woroch (1990) find conditions when the core fails to exist in much simpler network structures.

5.3.2. Non-co-operative approaches

An alternative formulation makes explicit the selection of strategies chosen by individual competitors. In this non-co-operative approach firms choose prices, qualities, and investment and also negotiate the terms of inter-carrier transactions. Compared with the co-operative approach, non-co-operative models have the potential to predict levels of these variables in equilibrium. They suffer from

some of the same infirmities as solutions to co-operative problems: multiple equilibria or no equilibrium at all.

An early example of the non-co-operative approach to local network competition, Economides and Woroch (1992), treated the case of the network access problem. This paper started from the 'rat-tail structure' described by Baumol (1983) where an entrant seeks to gain access to use of an incumbent's bottleneck facilities. Final and intermediate service prices are chosen non-co-operatively. They find that, among other results, equilibrium foreclosure of a non-integrated entrant depends on the extent to which the retail products of the two carriers are differentiated.

A different problem arises when two carriers initially compete for customers on equal terms and seek to interconnect to exchange two-way traffic. A typical approach to model competition for customers is to place them on a 'Hotelling' line with networks at endpoints[131]. The spatial differentiation now takes the form of an inherent preference by users for one network over the other. Demand for calls with all other users is assumed to be 'isotropic,' meaning they derive the same value from a call regardless of whom they connect to. In Economides, Lopomo and Woroch (1996a; 1996b), an incumbent network is distinguished by its ability to commit to pricing of retail and wholesale service prior to the arrival of the entrant. This assumption leads to the conclusion that the incumbent can structure originating and terminating prices so as to foreclose entrants from the market.

One other non-co-operative approach to local network competition treats the ILEC as a dominant firm and CLECs as fringe firms[132]. Here again the incumbent enjoys a strategic advantage in terms of its commitment to prices, while competitors take these prices as given.

5.4. *Entry barriers*

At various times in the above discussion, we pointed out different sources of barriers to entry into local exchange competition. Several of these barriers were natural, as with the strong scale and scope economies inherent in production of network services, and the huge sunk investments that are necessary for facilities-based supply. Other barriers are artificial when they are erected by regulations or legislation, as with the licensing and certification of competitive local carriers. Demand-created barriers are also prominent in the local services business as they are in many network industries. Positive feedback effects and user switching costs steer consumers away from new, small entrants. In fact, a principal concern of this section is the extent to which incumbents can strategically leverage these natural advantages to deter efficient entry into local network markets.

[131] Papers that adopt this modelling approach include Armstrong (1998b), Laffont, Rey and Tirole (1998a;1998b) and Economides, Lopomo and Woroch (1996a; 1996b).

[132] Examples of this approach include Abel (1999) and Plott and Wilkie (1995).

5.4.1. Natural barriers

In the cost section, we surveyed some of the econometric evidence on the presence of scale and scope economies in this industry. While this literature is inconclusive, it is reasonable that – for a sufficiently limited area such as a sparsely-populated residential neighbourhood – natural monopoly conditions prevail. Duplication of facilities is not the only way to inject competition into these markets, however. Entrants could simply share the use of existing facilities.

Scope economies can be the source of competitive advantage, provided they are supplemented with bundling strategies. Due to the costs and time involved in starting out, a local competitor cannot roll out the complete line of services from the beginning. An incumbent can defeat selective entry by bundling the service threatened by competitors with its protected services. Absent alternatives for these monopolised services, customers will prefer to buy from the incumbent as its marginal prices for the potentially competitive services are effectively zero.

5.4.2. Artificial barriers

Government intervention into local services markets is often responsible for barriers that competitors face. Many of these barriers are justified on efficiency grounds as in the case of patents and other intellectual property protection. The efficiency rationale for other restrictions on entry, such as licensing and certification of local competitors, are less apparent.

Access to certain essential resources is crucial to successful entry into local exchange markets, and some of these are controlled by government authorities. Wireline (and wireless) networks need access to rights of way, conduits, ducts, poles, and easements. Wireless networks need to locate their antennae, and more importantly, usually need rights to radio spectrum. Both carriers require telephone numbers if they wish to provide access service to end-users. The availability of this scarce resource is determined by property rights, and if phone numbers are allocated on a first-come basis, incumbents will have an advantage having claimed them over time. On the other hand, burdens are imposed on incumbents that entrants escape entirely, such as carrier-of-last-resort requirements and universal service obligations.

5.4.3. Strategic entry barriers

Strategic barriers are market conditions created by incumbent carriers that make entry more costly for prospective competitors, and that would not exist but for the threat of entry.

One of the best known means to erect a barrier is to make irreversible investments in durable assets[133]. Investment of this sort is unavoidable for wireline

[133] Dixit (1979).

networks as physical transmission paths tend to involve facilities that are very costly to redeploy. Compare that to a wireless network where the transmission path is the airwaves. In that case it is more a matter of whether expenditures to acquire rights to those airwaves are sunk, or whether the licenses are easily transferable at the market price. As a consequence of its sunk investment, an ILEC becomes a formidable competitor, willing to cut prices to a much lower level in the event of a price war with a competitor provided, of course, that the investment is observable by potential entrants. In a world of fast paced innovation, the advantage conferred by such investment is transitory, however. With each new generation of equipment, the incumbency advantage is at least partially neutralised.

Some embodiments of a first mover advantage are much less tangible than investment in switches and cables. By virtue of its history of serving the market, an incumbent has established a reputation with local customers whereas an entrant might be unknown, and hence, risky in the eyes of consumers[134]. Cable, long distance and electric utility companies have an advantage over *de novo* entrants into local exchange markets given their recognisable brand names and existing commercial relationships with potential customers in an area.

Loyalty to the incumbent may not derive just from the expectation of good service in the future. It may be caused by the costs anticipated by users should they choose to switch to a new supplier. It results in a reluctance of customers to switch to a new carrier even when it offers better price-quality package. As one example, upon switching to a new carrier, a user would have to choose a new phone number (if number portability was not required). Users would then have to notify all associates of the change, reprint new stationery and business cards, and so on. Even without these explicit costs, users display inertia in responding to alternative suppliers[135].

In other markets, incumbents have been known to supplement the loyalty their customers exhibit by signing them onto long-term contracts that discourage switching. Note that long term contracts and relationships with customers, especially large business customers, can cut the other direction as well: these contracts and relationships can limit the incumbent's range of actions, and render it more vulnerable to competitors[136]. In particular, to be technically compatible with its major customers, an ILEC may be reluctant to upgrade its network with the latest carrier equipment and technology.

An ILEC may sell services at volume discounts that are sufficiently large to make it unprofitable for an entrant to match assuming it cannot yet achieve

[134] An entrant could also be disadvantaged if it draws an unattractive sample of customers. This might occur if subscribers with bad payment records are diverted to new entrants. Customers who switch away from the incumbent may also be more likely to switch back adding to an entrant's high customer acquisition costs.

[135] See Knittel (1997) in the case of long distance service.

[136] See Christensen (1997) for a detailed discussion of how this occurs in the hard disk drive industry.

efficient long run scale. More explicit is the "market share discount" in which unit price falls as the percentage of service bought from the incumbent increases[137]. A multi-product firm also has the option of product bundling that can increase the sources of switching costs for a local service customer.

Another incumbent advantage stems from demand-side scale economies. Users place greater value on subscribing to a carrier's service the larger its customer base. Such network externalities encourage users to join the largest of the available networks, all else equal, and this in turn will tend to make the large network grow even larger relative to its competitors. This logic depends on the absence of interconnection among competing networks, for otherwise subscription to any one network would give a user access to all other users[138]. Interconnection neutralises any first mover advantage an incumbent might possess as a result of its larger customer base, provided that it does not re-create a pecuniary equivalent of network externalities by pricing traffic among its subscribes differently than traffic that travels between networks. In that case, once again a user will prefer the larger network to take advantage of the lower on-network rates[139].

An opportunity for strategic behaviour that is very ripe arises when ILECs are obliged to supply entrants with network services such as interconnection, collocation and unbundled network elements. By pricing these carrier services above cost, effectively selling the same services to its downstream affiliate at cost, the incumbent executes a 'price squeeze.' Similarly, in choosing the quality of these services, the incumbent can choke off a threat to its markets by supplying substandard access service to the entrant. Different models have come to different conclusions as to whether these exclusionary actions could be part of an equilibrium,[140] and empirical evidence does not settle the issue. In a closely related wireless context, Reiffen, Schumann and Ward (2000) fail to find conclusive evidence that the local wireline carrier favours its affiliate over the non-wireline competition in U.S. cellular markets relative to the alternative hypothesis of efficiencies between the two operations.

To sum up, technological economies – either scale or scope economies – are effective in achieving and maintaining dominance in the local service supply. They are neither necessary nor sufficient, however, for single firm to prevail in an unregulated market. An early provider of local services can leverage properties of that market – such as long-term contracts, brand recognition, network externalities, and user switching costs – to solidify its market position. Incumbents have

[137] A variant is the 'loyalty discount' whereby customers receive a rebate when they buy all of their services from the same provider.

[138] Residential subscribers sought to join the largest local network during the era of dual local systems, although businesses were compelled to subscribe to all services. Neither of these would be a concern if the networks were interconnected, both with one another and with all long distance networks.

[139] MCI innovated this kind of pricing policy with its "Friends and Family" program.

[140] Economides (1998) finds that the quality squeeze will occur while Sibley and Weisman (1998) come to the opposite conclusion.

natural disadvantages in competing in local service markets as well. For instance, over time, they are likely to become more heavily unionised, pay higher wage rates and have more restrictive work rules than a firm new to the market. How the balance of advantages and disadvantages among incumbents and entrants plays out depends in large part on regulatory policies that attempt to equalise conditions between the two firms.

6. Empirical evidence on local network competition

As a product of the short history of local competition, empirical research into its causes and consequences is sparse and idiosyncratic. What exists can be partitioned into investigations of the various determinants of competitive entry and the measurements of the economic effects of competition.

Many studies have examined effects of opening local exchange to competition at the country level[141]. The high level of aggregation does not permit tests of hypotheses of microeconomic effects, much less the nature of the strategic interaction among firms. For this reason, the research surveyed below is distinguished by being market and firm level.

We begin with empirical models that attempt to explain the incidence of local competition, and in some cases its timing as well. Woroch (1992) estimated a probit model of incidence of facilities-based CAP entry into the 120 largest U.S. cities using a using an original panel dataset that recorded deployment of fibre ring networks over the post-Divestiture period 1984–1991. As might be expected, population density of market and favourable state treatment of bypass are strong attracters for CAPs whereas ILEC fibre investment tends to discourage entry.

Zolnierek, Eisner and Burton (2001) examine the incidence and extent of local exchange entry following passage of TA96. Measuring entry by the number of carriers issued number code blocks in each of 190 LATAs, they estimate a multinomial logit model of entry as a function of LATA characteristics in each of four years, 1996–1999. The results confirm that highly populated and urbanised LATAs are the likely targets of most competition, as are areas served by one of the RBOCs.

Turning to empirical studies of competitive effects of local competition, Hausman, Tardiff and Ware (1989) was an early investigation study of the impact of local competition for business services. They measured changes in business use of long distance access services caused by entry of Teleport, Inc., Manhattan Cable, and others into the New York City market. They found that connection to alternative carriers reduced usage of switched long distance services by New York Telephone's large business customers significantly.

Another early study does not directly examine local entry effects, but is close enough to deserve attention. Mathios and Rogers (1988) examined state allow-

[141] See Ros (1999) and Walsten (2001).

ance of entry into the intraLATA toll market. They found that a state ban on facilities-based entry and resellers increased the average price of an intraLATA toll call.

A more recent contribution to this literature, one that uses the same entry measure as Zolnierik, Eisner and Burton (2001) is by Koski and Majumdar (2000). They examine strategic responses of U.S. ILECs to contemporaneous competition (measured by the number of firms with numbering resources) over a five-year period before and after the TA96. Using a panel over firms and years, they do not uncover a relationship between incumbent access pricing and entry[142]. However, Koski and Majumdar do find that ILECs raise advertising and become more focused on core telephony operations in response to competitive entry.

In evaluating all of these studies, extra caution should be exercised when interpreting their results to convey the cause-effect relationships governing competitive entry. To begin with, incumbent actions may not be clearly strategic in nature. Deployment of advanced network infrastructure as a cost reducing action could be confounded with an attempt to deter further entry. Even entrants' intentions may not be apparent. Certification by state commission, leasing rights of way, acquisition of numbering resources, and even construction of local networks do not necessarily represent true competition. Actual competition occurs only when the competitor begins to deliver services. Many instances exist where carriers merely acquire an option on future entry, or possibly pull out after an initial foray into a market.

To better expose the relationship between actions of incumbents and entrants, it is necessary to take full advantage of the inter-temporal dimension of the panel datasets. Toivanen and Waterson (2000) provide an example of this approach in a completely different industry. A similar approach was applied in Woroch (2000) to local exchange competition. That paper models ILEC and CLEC deployment of urban fibre rings in the U.S. over 1984–1992. Allowing for different lagged relationships, it is found that incumbents and entrants tend to match each other's deployments: entry triggers ILEC investment, and ILEC investment tends to invite competition.

7. Wireless local competition

This section examines the extent of current competition offered by various wireless technologies and the near-term prospects for this competition. Wireless service comes in two varieties: fixed and mobile. Fixed wireless is provided by a dedicated radio path linking a customer to the network facilities. Mobile wireless also establishes a radio link but here the customer can be anywhere in the serving territory, including travelling at fairly high ground speeds.

[142] Similarly, looking at rate levels charged by ILECs for DS3 special access services, Woroch (1992) fails to find a statistical dependence on the incidence of CAP entry.

The origins of wireless communications technology was as a means to improve the delivery of safety and emergency services and to facilitate communication on the battlefield. Commercial application of these technologies has achieved staggering success world-wide, as is clear from data on penetration and usage of these services. The ITU estimates approximately 720 million wireless lines world-wide as of the end of 2000 compared to 992 million main lines, and expressed in terms of population, there were 11.89 mobile lines per 100 population compared to 16.32 landlines[143]. Despite the lead of fixed line, the gap with wireless is closing rapidly: year on year growth in subscribers over the five-year period 1995–2000 was about 50 percent for mobile wireless but only 7 percent for fixed line. For this reason, the ITU projects equality for the two types of lines by about 2003. The popularity of fixed wireless has been much less impressive than mobile wireless. Only in developing countries has fixed wireless penetrated the residential sector to any extent.

Our concern, however, is not with the success of wireless technologies to achieve high penetration levels, or even with competition among wireless providers, but rather whether wireless technology industry constrains the pricing of incumbent (and entrant) wireline local service carriers. Does wireless service constitute a sufficiently close substitute to local wireline such that residential and business customers choose to use their wireless phones more often than their stationary phones, or even to replace their wireline phones with wireless alternatives? Are wireless providers capable of, and likely to, offer their services at reasonable prices over a wide area and with sufficient capacity to handle all voice and data traffic that currently travels on the PSTN?

Before turning to these questions, I give a brief, non-technical description of the wireless technologies with special attention to how they compare with the wireline alternatives, and then summarise the state of competition in the wireless submarkets[144].

7.1. Wireless communications technologies

7.1.1. Mobile wireless service

The first mobile wireless technology – Advanced Mobile Phone Service (AMPS) – was developed at Bell Laboratories in 1947. This system made an analogue radio connection between a transmitter tower and a user's handset. Frequencies were reused by partitioning a region into cells with a base station near the centre of each cell. The technology provided for "hand off" of a call as the user passed from one cell to another, even at highway speeds. The earliest commercial cellular phones were bolted into automobiles, and some time passed before transportables were introduced or today's miniature handsets appeared.

[143] ITU Telecommunications Indicators (2001).

[144] For more details see the Hausman chapter in this *Handbook*.

The first commercial cellular mobile service was launched by NTT in Tokyo in 1979. It was not until 1981 that the FCC decided to structure the U.S. cellular industry as a duopoly by creating two franchisees for each of 306 metropolitan areas and 428 rural areas. One license went to the local wireline incumbent serving each area while a second was awarded to a non-wireline independent carrier.

Cellular debuted in the U.S. when Illinois Bell first offered service in Chicago in October 1983. The following year Washington, D.C. became the first metropolitan market to offer users the choice of two cellular providers. During these early days, cellular coverage was spotty in large part because franchisees had not yet built out their networks. Invariably the wireline franchisee was ahead in the race to build the initial cellular network. To prevent the wireline franchise from dominating a market, the FCC required it to resell cellular network services to the non-wireline franchisee while it was still building out its network.

The technical quality of early systems was not good by today's standards due to rudimentary transmission equipment and propagation problems of the AMPS system. In addition to the poor quality of service, cellular service was very expensive across the board: handset equipment cost, activation and monthly subscription fees, and airtime charges. For these reasons it is little wonder users did not view cellular as a substitute for wireline service for many years to come.

Today, besides analogue cellular service, digital cellular is the leading technology with Personal Communications Services (PCS) widespread in the U.S. and Groupe Speciale Mobile (GSM) the standard throughout the rest of the world[145]. Paging services are also commonplace with two-way paging now beginning to appear in significant numbers. Less common mobile wireless technologies include Enhanced Mobile Radio Services (EMRS) such as the service offered by Nextel and satellite mobile phone systems such as the one Globalstar is deploying.

Modern mobile wireless technologies have made huge strides, both in terms of the technical quality of voice transmission and expanded vertical features. The second generation of mobile wireless was digital. Besides improved clarity, digital transmission greatly increased the carrying capacity of the congested frequency bands allocated to these services. Digitalisation allowed for coding of signals to prevent eavesdropping, a risk that remains a serious drawback for any analogue service. The Code Division Multiplexing Access (CDMA) protocol, originally developed for secure battlefield communication, makes signals virtually unbreakable. The new digital standards also added many vertical features that were bundled with voice telephony: paging, custom-calling features, voice mail, and now two-way email and web browsing.

[145] There were other analogue cellular standards besides AMPS that were used in various countries including Total Access Communications System (TACS) in Europe, the Nordic Telephone System (NTS) in Scandinavian countries, C-450 in Germany, NTT and JTACS in Japan and Radio Telephone Mobile System (RTMS) in Italy. See National Research Council (1997, Table 1–3).

The third generation (3G) wireless – sometimes called Universal Mobile Telephone Service (UMTS) – promises to greatly expand the bandwidth. Whereas current analogue and digital cellular services have a top data rate of 14.4 kbps (with 9.6 kbps being more common), 3G promises speeds of 2 Mbps[146]. This advance will enable Internet services over mobile phones comparable to those possible over a high-speed copper local loop and hybrid fibre-coaxial cable – with the added benefit of mobility.

Significantly, the newer digital technologies adopted smaller cell sizes which require reduced power levels[147]. This, in turn, reduced power requirements of handsets, making possible smaller batteries and longer talk times. Scale economies and the steady advances in design of digital signal processing (DSP) chips continue to drive down the costs of handsets and transmitter equipment. The smaller cells also made handoff more frequent, increasing the software and processing necessary to maintain service at highway speeds. The result was greater capital outlays required to build and interconnect the cell sites, raising the cost of deploying micro-cellular technology in sparely populated regions.

7.1.2. Fixed wireless service

This radio transmission technology replaces the copper loop, a coaxial drop or fibre drop with a high-frequency radio link, and for this reason is often called a wireless local loop (WLL). Economically, a radio link becomes more economical relative to these wireline alternatives the greater the distance between the user and the network switch. For this reason, it is not surprising that a very early fixed wireless service, BETRS (basic exchange telephone radio service), was deployed by local carriers to reach remote residential customers mainly living in rural areas[148]. More recently, higher frequencies have been developed to provide high-capacity point-to-point and point-to-multi-point connections. Multi-point multi-channel distribution system (MMDS) operates in the 24 GHz band. WinStar is an advocate of this technology as are several long distance companies seeking a wireless entry into local markets[149]. A second technology, Local Multi-point Distribution System (LMDS), is located in the higher 38 GHz ranges and, because of its enormous carrying capacity, is often referred to as "wireless fibre." By and large, both MMDS and LMDS technologies require line of sight to be most effective, and are vulnerable to rain fade and interference from foliage.

[146] FCC, Fifth Report on Wireless Competition, Aug. 2000, p. 36, fn. 237. Note that 3G speeds drop to 144 kbps when the user is travelling at highway speeds.

[147] I will not discuss the specialised wireless networks such as wireless PBXs installed in buildings and campus-wide wireless LANs.

[148] United Utilities in Alaska has offered BETRS as a customer option as opposed to the carrier making the choice as a wireless last mile technology.

[149] In the U.S., these bands were earlier used to distribute multichannel video, a service that was dubbed "wireless cable."

Fixed wireless technologies such as MMDS and LMDS are often referred to as 'big stick' technologies because they deploy a single tall antenna to serve each local area. They are especially well suited to high volume, data intensive business customers – especially those located in edge cities too far to justify building a dedicated fibre spur off an urban ring. The high frequencies allow transmission of huge amounts of voice and data to an interexchange carrier, an Internet backbone, or the company's local branch offices. In this respect, providers compete head to head with ILEC business services and with CAPs.

7.2. Wireless services as wireline competitors

Whether wireless providers can and will constrain the behaviour of local wireline carriers requires a comparison of supply and demand conditions of the two services. While fixed wireless will be discussed, the focus will be on the competitive threat posed by mobile wireless.

On the supply side, wireless networks are quicker to build and less costly to maintain, the principal reason being that a large portion of the transmission path is just airwaves. Once a carrier has a license to use the spectrum in the area, no additional investment is necessary beyond the transmitter and receiver equipment at the ends of the communication link. On the other hand, the airwaves can be hostile toward electromagnetic transmission. Adverse climate and terrain, idiosyncratic propagation properties, and radio wave interference all tend to reduce the overall reliability of wireless networks relative to wireline systems.

Wireless nonetheless has some features that make it part of an attractive entry strategy for a competitive local exchange company. Compared to wireline build out, a wireless network has a negligible marginal cost per line which does not vary with distance to the user[150]. For this reason, wireless has a particular advantage over wireline in serving sparsely populated areas. Another feature of wireless technology is that its infrastructure tends to be modular so that the network provider can incrementally add capacity as demand for its service grows[151]. This is important, for example, to an entrant who will inevitably share the local market with the wireline incumbent for many years to come.

On the other hand, radio spectrum can be a source of diseconomies when it is partitioned among several carriers. When, for instance, a block of spectrum is equally divided between two carriers, the amount of idle spectrum will necessarily rise. The reason lies in the fact that there will be times when one carrier has reached its spectrum limit, but not the other. In absence of a means to share

[150] Much of wireless expense is fixed relative to number of subscribers and their usage. So while it has a very small marginal cost quite independent of population density (though dependent on population mobility), it may have a high average cost which is sensitive to density.

[151] As an example, a cellular mobile network can partition its cell sites and use directional antennae to squeeze more capacity out of the same frequency band should the need arise. For more detail on this strategy, see the chapter by Hausman in this *Handbook*.

spectrum across carriers on a spot basis, less service will be provided than if a single unified wireless carrier utilised this bandwidth.

Wireless technologies of all kinds face several entry barriers not all of which are technological. First and foremost, licenses to use the airwaves are essential to any wireless venture. When those frequencies are not already occupied by a current tenant, and when these rights are auctioned to the highest bidder, the winning bids may extract much of the profit available from the service. The spectrum may also come with use restrictions which limit its usefulness to the carrier. It was only recently, for example, that the FCC permitted two-way transmission by LMDS license holders. Another, more tangible essential resource required by wireless providers are rights to locate their transmitter towers and related facilities. Wireless carriers can meet stiff resistance from communities seeking to preserve an aesthetic skyline or simply to exploit a potential revenue opportunity. The economics of wireless technologies, however, results in significantly lower entry barriers than their fixed line cousins.

On the demand side, we are especially interested in the extent to which users are willing to substitute a wireless alternative for their wireline service. More precisely, if the price of local wireline service were to increase by a significant amount, would users switch over to wireless alternatives in large numbers, either by replacing their fixed line with a wireless phone, or by shifting usage to mobile phones? Effectively, we are interested in the "diversion ratio" between wireline and wireless services: the percentage of the users who leave the PSTN who will turn to wireless in response to a wireline price hike. In conducting this thought experiment, we need to take account of initial wireless prices since wireless will not inhibit an increase in wireline prices if wireless is already very expensive. Also, if imperfect competition in wireless markets leads carriers to raise their prices in response, then little migration can be expected.

To assess whether wireless is a substitute for wireline, we begin by examining the properties of the two services. In several respects wireless and wireline provide the same local services: both provide access to the PSTN for incoming and outgoing calls; both offer a similar array of vertical features such as custom calling features and voice mail; both provide access to the Internet as well as to each others base of customers (assuming full interconnection).

Nevertheless each technology excels in certain areas. Wireline systems – whether the PSTN's copper loops or cable TV's coaxial cable – deliver much higher bandwidth with greater reliability using current technologies. Mobility is the key differentiating characteristic of wireless. Mobile wireless makes users 'accessible' so that they can receive calls at any time and in any place in the serving area. With fixed line service, one must be near the phone to receive calls, although voice messaging helps to fill the gap. In this direction, mobile wireless also satisfies demands for 'expediency' in that the users can place calls immediately rather than waiting until they reach a wireline phone, such as a public payphone. On the other hand, mobile wireless is a personal service with the phone carried by a single

individual. In contrast, different members of a household can more easily share the wireline by virtue of occupying the same house.

It is also possible, in principle, that the two services are complementary, at least for some users. Intra-household communication will certainly be facilitated when both types of lines are available to household members. In that case household members can be reached as they roam about the local area. Businesses may realise the same kind of benefits by connecting itinerant members of project teams.

In the end it is an empirical issue whether, on net, wireline and wireless services are substitutes or complements. While per-line usage of the PSTN in the U.S. has reached a plateau, growth of mobile wireless usage remains strong. Of course, these trends were greatly assisted by the relative price changes between wireline and wireless services. For instance between December 1997 and October 2000, it is reported that U.S. prices for cellular telephone service *fell* by 27.0 percent whereas the index for local charges *rose* by 9.8 percent[152]. Over the 10-year period ending in 1999, the number of cellular lines grew 2,359 percent while wireline subscriptions grew just 28.9 percent[153].

Econometric modelling is needed to isolate the portion of these trends that is attributable to substitutability between the two services. Using a sample of U.S. households having at least one wireline phone, Ahmad, Ward and Woroch (2002) find preliminary evidence that households treat the two services as substitutes in terms of usage. Households may substitute mobile wireless for wireline for non-local calls. Indeed Ahmad, Ward and Woroch (2002) find higher wireless usage prices lead to higher wireline long distance usage.

A form of substitution that would have more sustained competitive effects would, of course, be replacement of wireline with wireless service. In fact in most countries, growth of wireless lines exceeds landline, and in some (e.g., Norway, Korea, Japan) wireless *total* lines exceeds landline total[154]. While there is anecdotal evidence that some households – typically a single young adult – are relying exclusively on wireless service, the numbers in the population are small. Sung, Kim and Lee (2000) observe that, in Korea, not only have wireless sales overtaken wireline, but users are also disconnecting their traditional phones[155].

Another form of substitution occurs when new subscribers (usually a result of in-migration or new household formation) opt for wireless access rather than

[152] Bureau of Labor Statistics, Consumer Price Index Detailed Report, October 2000, Table 25. See also the chapter by Hausman in this *Handbook*.

[153] CTIA (2001) and FCC Trends in Telephone Service (2000).

[154] ITU (2001b) counts 35 countries as of mid-2001 in which mobile phones had overtaken fixed lines.

[155] In one study, it has been estimated that between 1990 and 1996, 60,000 landlines were displaced by wireless lines in Nordic countries (OECD, 1995). More recently, the ITU (2001b) notes that in the last ten years, household penetration of fixed line in Finland has dropped from 94 percent down to 83, during which time mobile has increase from 7 to 60 percent.

fixed service. The percentage of individuals who depend exclusively on mobile wireless service has been estimated to be 3 percent in the U.S[156]. and 6 percent in the U.K[157]. Alternatively, as a household's demand for communications increases, it may choose to meet that demand with mobile wireless. The FCC recently reported that survey results show that 12 percent of households chose mobile service rather fixed line when adding a second line[158].

It is important to emphasise the role that pricing plays in determining demand for mobile wireless service and its substitution for wireline. Typically, mobile service is more costly than wireline, though that difference is shrinking. As of December 2000, the CTIA reports that the average cellular bill was $45.27 while the FCC estimates the typical local wireline bill to be $34 in that same year[159]. Of course, these figures must be compared understanding that the usage level of the typical line could be vastly different for the two services. For instance, since cellular and PCS services in the U.S. usually charge by the minute – whether incoming or outgoing – usage can be expected to be less for the same user. Until a calling-party pays (CPP) system is implemented, and until measured local service becomes prevalent, this pricing regime will exert a drag on wireless usage. Prepaid mobile service would counteract the effects of high relative usage prices for wireless. So far, however, prepaid service has not been as popular in the U.S. as it has elsewhere.

7.3. Structure of the wireless industry

Supply and demand substitutability is a necessary condition for competition between wireline and wireless services but alone it is not sufficient. Wireless services may be priced very high relative to wireline substitutes, due to imperfect competition, discouraging users from making the switch. The first decade of the U.S. cellular industry illustrates this situation. Structured as a duopoly, with limited FCC and state regulation, prices would tend to be higher than under unfettered competition[160]. The limited spectrum allocated to the service relative to the requirements of the original AMPS technology contributed to the lack of competition. At the end of 2000, the HHI of the U.S. cellular and PCS industry was 1,564 measured on a nation-wide basis in terms of subscribers[161].

[156] FCC, Sixth Report of Wireless Competition (2001).
[157] OFTEL (2000b).
[158] FCC (2001), op.cit..
[159] FCC Trends (2000, Table 3.2).
[160] In fact, Hausman maintains that a number of states regulated mobile service and that regulation resulted in *higher* prices than in those states that had no regulation at all. See, the chapter by Hausman in this *Handbook*.
[161] Computed from subscriber totals reported by the largest 25 mobile wireless providers in the FCC's Sixth Report on Wireless Competition (2001, Table 3, p. C-4). These data are on a national level; taking an average of HHIs computed for individual MSA and RSA markets likely will produce higher levels of concentration.

This represents a decrease from an HHI of 1,846 from one year earlier. Using data from the early 1990s Parker and Röller (1997) find econometric evidence that, prior to the introduction of PCS, the cellular duopoly in the U.S. was imperfectly competitive, and trace it to multimarket contact and cross ownership among cellular providers. Ruiz (1994) and Fullerton (1998) reach more mixed conclusions when testing for various kinds of collusive behaviour.

Another reason why mobile wireless markets might not achieve competitive outcomes is ownership of one (or more) of the wireless carriers by the incumbent wireline company. A vertically integrated ILEC could execute a price squeeze on its wireless competitors by charging high rates to terminate their traffic on the fixed network. The ILEC could also subsidise its wireless operations by shifting costs to its wireline side assuming that business operates under cost-based regulation.

With the launch of PCS in 1996, the U.S. wireless industry was quickly transformed from an industry of isolated, geographic duopolies to one populated by several wireless oligopolists. Many of these new carriers had coast-to-coast footprints as a result of large mergers among wireless carriers and the availability of national PCS licenses. Today in the U.S. there are six nation-wide wireless carriers[162].

Competition among mobile carriers drives down wireless prices, and raises service quality for the user, and in the process, wireless becomes a more attractive alternative to wireline service. Besides the advent of PCS, intra-wireless competition has intensified in recent years for several other reasons. New digital radio technologies have expanded the array of possible wireless services. PCS in the U.S. and GSM elsewhere are examples. The FCC reports that, by the end of 2000, nearly 91 percent of the U.S. population had available three or more mobile wireless providers, and nearly 75 percent had five or more[163].

Outside the U.S., privatisation of state-owned wireless carriers and the opening of existing wireless markets to entry by private carriers intensified competition in these markets[164]. In markets such as Europe, CPP and pre-pay systems and the prevalence of measured fixed service added to the growth of wireless service relative to wireline.

7.4. An assessment of the wireless threat

We can expect that, in time, the threat posed by wireless will grow as competition among mobile providers further drives down wireless rates and as deregulation

[162] AT&T, Sprint, Nextel, Verizon, VoiceStream, and a joint venture between BellSouth and SBC, Cingular Wireless.

[163] FCC, Sixth Report on Wireless Competition (2001, Table 4).

[164] OECD (2001) reports that presently all of the mobile communications industries of its 29 member countries are competitive when 23 of them were organized as monopolies a decade earlier.

continues to rebalance local basic service charges[165]. Technological advances will continue to close the quality and bandwidth gaps between the two technologies. Relentless build out of wireless networks has enveloped ever-larger serving areas. In the very near future, wireless will overtake wireline in both access lines and usage in many of the major developed countries of the world.

Despite the pressure on price and quality, wireless is not likely to supplant the wireline network anytime soon. The PSTN offers considerable advantages. Foremost is the fact that the wireline network is already built and ubiquitous. On the data front, wireline is likely to maintain its lead as the two technologies will continue their cat-and-mouse race to ever-greater bandwidth.

More likely, wireless will fill the geographical and product gaps left open by the wireline network. New wireless technologies on the commercial horizon promise to do exactly this. 'Ultra wide-band wireless' technologies take advantage of under used frequencies scattered throughout the radio spectrum to deliver data. Other emerging wireless technologies do not use the electromagnetic spectrum at all. 'Free space optics,' for instance, transports information on low-power laser beams between two points within line of sight.

Many of these technologies are speculative and some are sure to fail. One technology that was touted as highly promising and attracted enormous financial backing a few years ago was satellite mobile phone. Launching dozens of low earth orbit and middle earth orbit satellites which function as cell sites streaming across the sky, these systems promised to reach the most remote regions on the globe. The systems were very costly to build, operate and maintain, and consequently resulted in very costly handsets and high per-minute rates. In the end, several well-financed projects lost billions of dollars and ended in bankruptcy – including Motorola's Iridium and ICO Global Communications. Furthermore, satellite Internet access services such as the one offered by Teledesic use telephone dialup access for the uplink portion of the connection[166].

This is an important feature of all wireless technologies: while they can technically substitute for wireline service, they will not operate completely independent of the PSTN for the foreseeable future. Even a connection that is wireless at both ends must travel through an earth-bound switch. Effectively fixed and mobile wireless technologies append a "radio tail" to a wireline network. To be of value to prospective users, a wireless network must physically interconnect with the PSTN so that the user can reach land-line users. In that case the wireless providers must reach agreement with wireline networks to mutually terminate traffic at affordable rates. Progress has been made in this direction. In the U.S., law requires wireline common carriers to interconnect with commercial mobile radio services[167]. The WTO's Basic Agreement on

[165] In fact, Federal law forbids taking account of CRMS when determining the extent of competition for wireline incumbents. Omnibus Budget Reconciliation Act of 1993, 47 U.S.C. 332.

[166] Two-way satellite high-speed data networks are being deployed however.

[167] Omnibus Budget Reconciliation Act of 1993, 47 U.S.C. 332(c)(1)(B).

Telecommunications also requires interconnection at non-discriminatory rates. Note that, until the TA96, cellular networks paid to receive traffic from ILECs as well as paying them to terminate mobile traffic. Another important obstacle, one that is crucial to entry into the residential market, is the absence of number portability between wireline and wireless systems, as well as among wireless carriers.

8. The future of local competition

Where market forces are allowed to operate, history shows that the local exchange industry tends to swing between monopoly and competition. Over the years, New York City illustrated this pattern in high relief. The current wave of competition in the U.S. and elsewhere is not unique, though it has proved to be more substantial, and it has proved to be more sustained than previous episodes.

While monopoly over local network markets has been the rule rather than the exception, competitive pressure on local services markets has been incessant. Often the assaults are indirect, attacking a narrow niche that is either outside the purview of regulators or beyond the principal interests of incumbent providers[168]. On occasion, these forays establish a beachhead that later expands to compete with the incumbent's core markets. This occurred when CAPs began as carriers' carriers and then gradually migrated into the delivery of switched services to homes and businesses.

Invariably at the source of successful entry is some technological advance that enables a new service or is a new way of delivering an existing service. A recent example is Internet telephony. IP telephony was originally applied for international calling, but now it is emerging as an alternative platform for local calling, spurred on by the phenomenal growth of the Internet.

Market forces that drove the industry toward high concentration in the past nevertheless remain strong and pervasive today. Scale and scope economies deriving from network structure have not vanished. Network externalities that reward first movers and large incumbents are less prominent in mature telecommunications markets where penetration is nearly complete and all major networks are interconnected. Rather, in today's more competitive environment, the focus has turned to 'ownership' of retail customers. A product of supplier reputation and user switching costs, this demand side effect also works to the advantage of large-scale producers.

Policies that seek to inject competition into local network markets by sharing incumbent networks with rivals seek to have the best of both worlds. Unbundling of network services and resale of retail services preserve the benefits of unified

[168] Several examples of this pattern exist in data communications services. See Leo and Huber (1997).

production of network services while at the same time facilitating competition for retail services. The same is true for policies of structural separation of network services.

From all indications these approaches have failed to take full account of the transaction costs incurred when sharing facilities and resources among competing providers. These transactions have triggered much haggling, contracting and monitoring, and in the worst cases, litigation and enforcement. The extent of the transaction costs, broadly interpreted, generated by implementation of the TA96 was clearly an unpleasant revelation for its framers.

An important source of these costs can be traced back to the misalignment of incentives of incumbent providers and new entrants. In the end, unbundling is an unnatural act for a vertically integrated provider. It is no surprise, therefore, that unbundling was virtually unknown prior to the recent opening the local exchange. Instead we saw fierce battles over interconnection of competing networks dating from the earliest days of the industry. Realistically, a goal of perfect interconnection, or the complete absence of discriminatory treatment of affiliated and unaffiliated partners, is unattainable. The embedded local networks we have today were optimized for exclusive use by a monopoly carrier, not for wholesale supply of unbundled elements or other network services.

The truth is, many geographic areas and customer segments – especially small markets with low population density, or customer groups with highly specialized service demands – are efficiently organized as monopolies. As a consequence, a prescription of ubiquitous competition may be no less harmful to social welfare than an integrated monopoly in each and every market.

In addition, service-based competition is inherently limited because competitors are restricted by the price, service and technology choices of the infrastructure owner. At best, over the long run, it offers a stepping stone to competitors on their way to building access networks of their own. Facilities-based competitors do not suffer from these same infirmities, and because of the durability of their investments, entry of this kind is more likely to have a sustained impact.

When existing networks are redeployed to provide local service (e.g., cable telephony and electric powerline systems), facilities-based entry can be relatively quick. These alternatives do not avoid the time and expense of negotiating interconnection agreements with incumbents, nor the risk of shifts in regulatory policy or legal rulings toward this kind of competition. But the incremental expense of entering with facilities, as well as reductions in associated sunk investment, make this a particularly effective and attractive competitor to incumbent carriers.

Experiments with open competition underway around the world offer tests of the relative merits of infrastructure and service competition. Comparison of the experience in the U.S. and U.K. is a case in point. Whereas the U.K. has favoured facilities-based entry ever since privatisation of BT, the U.S. was a leader in implementing network unbundling and resale. By the end of 1999, fixed line

competitors were reported to have achieved a 15.4 percent share of access lines in the U.K., three times the 5.44 percent penetration achieved in the U.S. the majority of which is resold local loops[169]. Of course, market and institutional conditions differ significantly between the two countries, as did the development of incumbent and alternative networks at the time when local markets were opened to competition. Yet the similarities between the two countries were close enough to make the comparison instructive.

Facilities based competition faces its own obstacles quite aside from its enormous capital requirements. As with any network, facilities-based entrants must locate their equipment and links over land, under ground and through the air. Acquiring rights of way is essential, if often tedious, as when gaining access to building tops and riser space. Municipal authorities can be stingy with their public resources, and often tax the new providers.

To tap the benefits of infrastructure competition, policy makers must resolve several difficult issues. Arguably less challenging than implementing service competition, regulators nevertheless must strive to extend symmetric treatment to different carriers, different regions and different services. Efficient policy toward incumbents and new entrants is particularly nettlesome, and the problem of incremental infrastructure investment poses sticky issues. Opening these facilities to competition will diminish incentives to build it in the first place, but if not, then competitors will necessarily stand at a competitive disadvantage. A good example of this is the 'next generation network' (NGN) that has been predicted for some time. This all-optical, all-packet network is intended to supplant the ageing PSTN but, given their position in the market, ILECs will build at least a portion of the NGN in all likelihood.

It seems inevitable that, in the end, any initiative to open local networks to competition must undergo a long, arduous transition period, especially coming after decades of regulated monopoly or state ownership. During this time, firms must learn how to compete, whether they are entrants seeking to break into local markets, or incumbents responding to the competitive threats. Consumers are coping with a wide array of providers and the plethora of services options and new technologies. Regulators must grope their way toward means to aid entry by new competitors without destroying investment incentives of incumbent carriers. At this time, too little history is available to draw inferences about the effects of alternative policies on local network markets. Time is needed before it is known how this process will operate and which policies are superior. Meanwhile, it is important to let the experiment run its natural course.

[169] OECD (2001 Table 2.4). Recent survey results from the U.K. in early 2000 reveal that no fewer than 22 percent of households obtain a fixed line service exclusively from a competitive supplier, invariably a cable operator, with 28 percent taking some fixed line service from a BT competitor. See OFTEL (2000d, Figure 3a).

References

Abel, J., 1999, Entry in regulated monopoly markets: The development of a competitive fringe in the local telephone industry, Columbus, OH: National Regulatory Research Institute.

Abel, J. and M. Clements, 1998, A time series and cross-sectional classification of state regulatory policy adopted for local exchange companies: Divestiture to present, 1984–1998, Columbus, OH: National Regulatory Research Institute 98-25.

Ahmad, F., M. Ward and G. Woroch, 2002, Are fixed and mobile services good substitutes? Do wireless and wireline services compete? presented at the PVRC Conference, "Competition in Wireless" University of Florida, Gainsville, FL.

Anderson, R., T. Bikson, S. A. Law, B. Mitchell, 1995, Universal access to email: Feasibility and social implications, Rand Report, Centre for Information Revolution Analyses.

Armstrong, M., 1998a, Local competition in UK telecommunications, in M. Beesley, ed., Regulating Utilities: Understanding the Issues, London: Institute of Economic Affairs.

Armstrong, M., 1998b, Network interconnection in telecommunications, Economic Journal, 103, 545–564.

Armstrong, M., 2002, The theory of interconnection and access pricing, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: North-Holland, 295–384.

Barnett, W. and G. Carroll, 1993, How institutional constraints affected the organisation of early U.S. telephony, Journal of Law, Economics & Organisation, 9, 98–126.

Baumol, W., 1977, On the proper cost tests for natural monopoly in a multi-product industry, American Economic Review, 67, 809–22.

Baumol, W., 1983, Some subtle pricing issues in railroad regulation, International Journal of Transport Economic, 10, 341–355.

Baumol, W. and J. G. Sidak, 1994, Toward competition in local telephony, Cambridge, MA: MIT Press.

Bittlingmayer, G., 1990, Efficiency and entry in a simple airline network, International Journal of Industrial Organisation, 8, 245–57.

Bornholz, R. and D. Evans, 1983, The early history of competition in the telephone industry, in D. Evans, ed., Breaking Up Bell: Essays on Industrial Organisation and Regulation, Amsterdam: North-Holland.

Brock, G., 1981, The telecommunications industry: The dynamics of market structure, Boston: Harvard University Press.

Brock, G. W., 1994, Telecommunications policy for the information age: From monopoly to competition, Cambridge, MA: Harvard University Press.

Cellular Telephone Industry Association, 2001, Semi-annual mobile telephone industry survey, Washington, D.C.: CTIA.

Christensen, C., 1997, The innovator's dilemma: When technologies cause great firms to fail, Boston: Harvard Business School Press.

Crandall, R. W., 2001, An assessment of the competitive local exchange carriers five years after the passage of the Telecommunications Act, a report for the USTA by Criterion Economics, LLC.

Crandall, R. and T. Hazlett, 2000, Telecommunications policy reform in the United States and Canada, working paper, American Enterprise Institute.

Deutsche Bank, A. Brown, 2001, The state of the network: Technology, capital and regulation are key attributes in today's telecommunications networks, Gary P. Jacobi and Eric Melloul.

Dixit, A., 1979, A model of duopoly suggesting a theory of entry barriers, Bell Journal of Economics, 10, 20–32.

Economides, N. and G. Woroch, 1992, Benefits and pitfalls of network interconnection, Discussion Paper No. EC-92-31, Stern School of Business, New York University.

Economides, N., G. Lopomo and G. Woroch, 1996a, Strategic commitments and the principle of reciprocity in interconnection pricing, Discussion Paper EC-96-13, Stern School of Business, New York University.

Economides, N., G. Lopomo and G. Woroch, 1996b, Regulatory pricing policies to neutralize network dominance, Industrial and Corporate Change, 5, 1013–1028.

Economides, N., 1998, The incentive for non-price discrimination by an input monopolist, International Journal of Industrial Organisation, 16, 271–284.

Evans, D. and J. Heckman, 1983, Multi-product cost function estimates and natural monopoly tests for the Bell System, in D. Evans, ed., Breaking Up Bell: Essays on Industrial Organisation and Regulation, Amsterdam: North-Holland.

Fagen, M. D., 1975, A history of engineering and science in the bell system: The early years 1875–1925, Bell Telephone Laboratories.

Faulhaber, G., 1975, Cross-subsidisation: Pricing in public enterprises, American Economic Review, 65, 966–977.

Federal Communications Commission, 1939, Investigation of the telephone industry in the United States, Report of the U.S. Federal Communications Commission on The Investigation of the Telephone Industry in the United States Made Pursuant to Public Resolution No. 8, 74th Congress, Washington, D.C.: U.S. Government Printing Office.

Federal Communications Commission (annual), Statistics of Communications Common Carriers, Washington, D.C.: Industry Analysis Division, Common Carrier Bureau.

Federal Communications Commission (annual), Trends in Telephone Service, Washington, D.C.: Industry Analysis Division, Common Carrier Bureau.

Federal Communications Commission, 1999, Memorandum Opinion and Order, Application of Bell Atlantic of New York for Authorization under Sect. 271 of the Communications Act to Provide In Region InterLATA Service in the State of New York, CC Docket 99-295.

Federal Communications Commission, 2000a, Fifth Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, Wireless Services Bureau.

Federal Communications Commission, 2000b, Local telephone competition at the new millennium, Washington, D.C.: Industry Analysis Division, Common Carrier Bureau.

Federal Communications Commission, 2000c, Local telephone competition: Status as of June 30, 2000, Washington, D.C.: Industry Analysis Division, Common Carrier Bureau.

Federal Communications Commission, 2001a, Local telephone competition: Status as of December 31, 2000, Washington, D.C.: Industry Analysis Division, Common Carrier Bureau.

Federal Communications Commission, 2001b, Sixth Annual Report and Analysis of Competitive Market Conditions With Respect to Commercial Mobile Services, Washington, D.C.: Wireless Services Bureau.

Federal Trade Commission and Department of Justice, Horizontal Merger Guidelines, Washington, D.C.: Government Printing Office.

Friedlander, A., 1995, Natural monopoly and universal service: Telephones and telegraphs in the U.S. communications infrastructure, 1837–1940, Reston, VA: Corporation for National Research Initiatives.

Fullerton, H., 1998, Duopoly and competition: The case of American cellular telephone, Telecommunications Policy, 22, 593–607.

Fuss, M. and L. Waverman, M., 2001, Econometric cost models, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Amsterdam: Elsevier Science, BV, forthcoming.

Gabel, D. and J. Nix, 1993, AT&T's strategic response to competition: Why not preempt entry? Journal of Economic History, 53, 377–387.

Gabel, D., 1994, Competition in a network industry: The telephone industry, 1894–1910, Journal of Economic History, 54, 543–572.
Gabel, D. and M. Kennet, 1994, Economies of scope in the local telephone exchange market, Journal of Regulatory Economics, 6, 381–398.
Gabel, D. and D. Weisman, 1996, Historical perspectives on competition and interconnection between local exchange companies: the United States, 1894–1914, in D. Gabel and D. Weiman, eds., Opening Networks to Competition: The Regulation and Pricing of Access, Boston: Kluwer Publishing.
Gabel, R., 1969, The early competitive era in telephone communication, 1893–1920, Law and Contemporary Problems, 34, 340–59.
Geroski, P., 1992, Market dynamics and entry, Basil Blackwell: Oxford.
Gort, M. and N. Sung, 1999, Competition and productivity growth: The case of the U.S. telephone industry, Economic Inquiry, 37, 678–691.
Hausman, J., 2002, Mobile telephone, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science, B. V., 563–604.
Hausman, J., T. Tardiff and H. Ware, 1989, Competition in telecommunications for large users in New York, in Telecommunications in a Competitive Environment, White Plains, N.Y.: National Economic Research Associates.
International Telecommunications Union, 2001a, Telecommunication indicators – basic and cellular, Geneva.
International Telecommunications Union, 2001b, ITU Telecommunication Indicators Update.
Huber, P., 1987, The geodesic network: 1987 Report on competition in the telephone industry, U.S. Department of Justice, Antitrust Division, Washington, D.C.: U.S. Government Printing Office.
Katz, M. and G. Woroch, 1997, Introduction to special issue on telecommunications, Industrial and Corporate Change, 6, 701–718.
Knittel, C., 1997, Interstate long distance rates: Search costs, switching costs and market power, Review of Industrial Organisation, 12: 4, 519–536.
Koski, H. and S. Majumdar, 2000, Paragons of virtue? Patterns of entry and the strategies of incumbents in the U.S. local telecommunications industry, Mimeo, University of London: Imperial College of Science, Technology and Medicine.
Kridel, D. and L. Taylor, 1993, The demand for commodity packages: The case of telephone custom calling features, Review of Economics and Statistics, 362–368.
Laffont, J.-J., P. Rey and J. Tirole, 1998a, Network competition I: Overview and non-discriminatory pricing, The RAND Journal of Economics, 29, 1–37.
Laffont, J.-J., P. Rey and J. Tirole, 1998b, Network competition II: Discriminatory pricing, The RAND Journal of Economics, 29, 38–56.
Laffont, J.-J. and J. Tirole, 2000, Competition in telecommunications, Cambridge, MA: MIT Press.
Leo, E. and P. Huber, 1997, The incidental, accidental deregulation of data and everything else, Industrial and Corperate Change, 6, 807–828.
Mathios, A. and R. Rogers, 1988, The impact of state price and entry regulation on intrastate long distance telephone rates, Federal Trade Commission Report.
Melo, J.R., 1998, Chile, in E. Noam, ed., Telecommunications in Latin America, New York: Oxford University Press.
Merchants Association of New York, 1905, Inquiry into telephone services and rates in New York City.
Mikolas, M., 1990, RBOC Strategies for local loop reliability, Business Communications Review, 10, 48–53.
Mitchell, B., 1990, Incremental costs of telephone access and local use, Rand Report R-3909-ICTF.
Mueller, M., 1997, Universal service: Competition, interconnection and monopoly in the making of the American telephone system, Cambridge, MA: MIT Press.

National Research Council, 1997, The evolution of untethered communications, Washington, D.C.: National Academy Press.
New York Public Service Commission, 2001, Analysis of local exchange service competition in New York State: Reflecting company reported data and statistics as of December 31, 2000, Albany, NY.
Noam, E., 2002, Interconnection practices, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science, B.V., 385–421.
OECD, 1995, Mobile and PSTN communication services: Competition or complementarity? OECD/GD (95) 96, Paris: OECD.
OECD, 1999, Communications outlook, 1999, Paris: OECD.
OECD, 2001, Communications outlook 2001, Paris: OECD.
OFTEL, 1999, Access to bandwidth: delivering competition for the information age, statement issued by the Director General of Telecommunications, London.
OFTEL, 2000a, Mobiles – Barriers to switching, and quality of service issues – Summary of consumer research conducted for OFTEL by Ipsos-RSL.
OFTEL, 2000b, Consumer awareness and use of competition in the residential fixed line market, London.
OFTEL, 2000c, Consumer switching behaviour in the telecoms market, London.
OFTEL, 2000d, Consumers' use of fixed telecoms services: Summary of OFTEL Residential Survey, London.
OFTEL, 2001, Market information: fixed update, London.
Palmer, K.,1992, A test for cross subsidies in local telephone rates: Do business customers subsidise residential customers? RAND Journal of Economics, 23, 415–431.
Panzar, J. C., and R. Willig, 1981, Economies of scope, American Economic Review, 71, 268–272.
Park, R. E., B. Wetzel and B. Mitchell, 1983, Price elasticities for local telephone calls, Econometrica, 51, 1699–1730.
Parker, P. M. and L. H. Röller, 1997, Collusive conduct in duopolies: Multimarket contact and cross-ownership in the mobile telephone industry, RAND Journal of Economics, 28, 304–22.
Perl, L., 1983, Residential demand for telephone service, 1983, prepared for the Central Service Organisation of the Bell Operating Companies by National Economic Research Associates, Inc., White Plains, NY.
Plott, C.R. and S. Wilkie, 1995, Local telephone exchanges, regulation and entry, California Institute of Technology, Division of the Humanities and Social Sciences, Working Paper 941.
Reiffen, D., L. Schumann and M. Ward, 2000, Discriminatory dealing with downstream competitors: Evidence from the cellular industry, Journal of Industrial Economics, 158, 253–286.
Rochester Telephone Co., 1993, Petition for approval of proposed restructuring plan, filed with the New York Public Service Commission.
Röller, L.-H., 1990, Proper quadratic cost functions with an application to the Bell system, Review of Economics and Statistics, 72, 202–210.
Ros, A., 1999, Does ownership or competition matter? The effects of telecommunications reform on network expansion and efficiency, Journal of Regulatory Economics, January, 15, 65–92.
Roycroft, T., 1998, A dynamic model of incumbent LEC response to entry under the terms of the Telecommunications Act of 1996, Journal of Regulatory Economics, 14, 211–228.
Ruhle, E.-O., 1999, 161 days of full competition: Some observations from the German market, in D. G. Loomis and L. D. Taylor, eds., The Future of the Telecommunications Industry: Forecasting and Demand Analysis, Boston: Kluwer Academic Publishers.
Ruiz, K. L., 1994, Regulatory effects in the U.S. Cellular telecommunications duopolies, Ph.D. Dissertation, Boston University.
Sappington, D. E. M., 2002, Price regulation, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elseveir Science B. V., 225–293.

Sharkey, W. W., 1982, The theory of natural monopoly, Cambridge: Cambridge University Press.
Sharkey, W., 1991, Supportability of network cost functions, Annals of Operations Research, 36, 1–16.
Sharkey, W. W., 1995, Network models in economics, in M.O. Ball et al., eds., Network Routing, Vol. 8, Handbooks in Operations Research and Management Science, Amsterdam: Elsevier Science.
Sharkey, W., 1999, The design of forward-looking cost models for local exchange telecommunications networks, Washington, D.C.: Federal Communications Commission, Working paper.
Sharkey, W. W., 2002, Representation of technology and production, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science, B. V., 179–222.
Shin, R. and J. Ying, 1992, Unnatural monopoly in local telephone, Rand Journal of Economics, 23, 171–83.
Shin, R. and M. Ward, 1993, Comments of the staff of the FTC on CC Docket 91-141, Expanded Interconnection with Local Telephone Company Facilities.
Sibley, D. and D. Weisman, 1998, Raising rivals' costs: the entry of an upstream monopolist into downstream markets, Mimeo.
Spulber, D., 1991, Regulation and markets, Cambridge, MA: MIT Press.
Standage, T., 1998, The Victorian Internet: The remarkable story of the telegraph and the nineteenth century's on-line pioneers, New York: Berkley Publishing.
Stehman, J. W., 1925, The financial history of the AT&T Company, Boston: Houghton-Mifflin.
Stone, A., 1997, How America got online: Politics, markets and the revolution of telecommunications, Armonk, NY: M.E. Sharpe.
Sung, N., C.-G. Kim and Y.-H. Lee, 2000, Is a POTS dispensable? Substitution effects between mobile and fixed telephones in Korea, paper presented at International Telecommunications Society biennial conference, Buenos Aires.
Taylor, L., 1999, Telecommunications demand analysis in transition: An overview of part I, in D. G. Loomis and L. D. Taylor, eds., The Future of the Telecommunications Industry: Forecasting and Demand Analysis, Boston: Kluwer Academic Publishers.
Taylor, L. D., 2001, Customer demand analysis, in M. E. Cave, S. K. Majumdar and I. Vogelsang, eds., Handbook of Telecommunications Economics, Volume 1, Amsterdam: Elsevier Science B.V., 97–142.
Toivanen, O. and M. Waterson, 2000, Market structure and entry: Where's the beef? Mimeo.
Tomlinson, R., 1995, The impact of local competition on network quality, in W. Lehr, ed., Quality and Reliability of Telecommunications Infrastructure, Mahwah, NJ: Lawrence Erlbaum Associates.
Tomlinson, R.G., 2000, Tele-revolution: Telephone competition at the speed of light (A history of the creation of a competitive local telephone industry: 1984–2000), Glastonbury, CT: Penobscot Press.
U.S. House of Representatives,1915, The postalisation of the telephone, Report of the Committee on Post Office and Post Roads.
Vogelsang, I. and B. Mitchell, 1997, Telecommunications competition: The last ten miles, Cambridge, MA and Washington, D.C.: MIT Press and AEI Press.
Vogelsang, I. and G. Woroch, 1998, Local telephone service: A complex dance of technology, regulation and competition, in L. Duetsch, ed., Industry Studies, Armonk, NY: M. E. Sharpe, 2nd Edition.
Wilcox, D., 1910, Municipal Franchises, Vol. 1, Rochester, NY: Gervaise Press.
Woroch, G., 1987, On pricing with lumpy investment: Two problems of Boiteux, University of Rochester, Department of Economics Working Paper.
Woroch, G., 1990, On the stability of efficient networks: Integration vs. fragmentation in communications and transportation, Mimeo.

Woroch, G., 1992, An empirical model of innovative entry and a test of strategic deterrence: An application to local access markets, paper presented at the Telecommunications Policy Research Conference.

Woroch, G., 1995, Reply affidavit filed in U.S. v. Western Electric Co., Inc. and AT&T Co., U.S. District Court For The District Of Columbia, Civil No. 82-0192 (HHG).

Woroch, G., 1996, Turning the cables: Economic and strategic analysis of cable entry into telecommunications, in E. Noam, ed., Globalism and Localism in Telecommunications, Amsterdam: Elsevier Science.

Woroch, G., 2000, Competition and investment in digital infrastructure, paper presented at International Telecommunications Society biennial conference, Buenos Aires.

Zolnierek, J., J. Eisner and E. Burton, 2001, An empirical examination of entry patterns in local telephone markets, Journal of Regulatory Economics, 19, 143–160.

SUBJECT INDEX