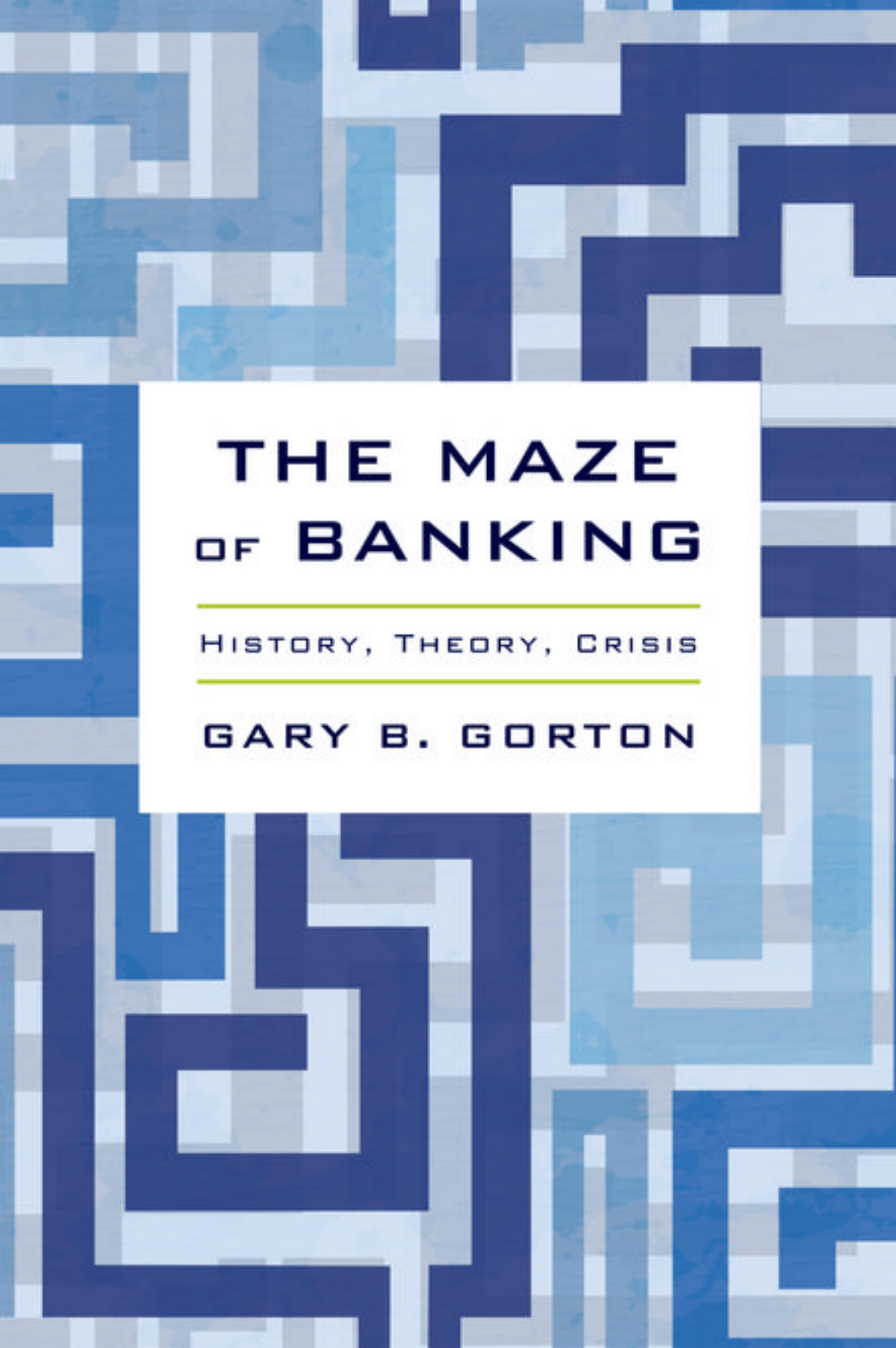

THE MAZE
OF BANKING
HISTORY, THEORY, CRISIS
GARY B. GORTON

The Maze of Banking

The Maze of Banking

History, Theory, Crisis

GARY B. GORTON

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of
Oxford. It furthers the University's objective of excellence in research,
scholarship, and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries.

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016,

CIP data is on file at the Library of Congress
ISBN 978–0–19–020483–9

9 8 7 6 5 4 3 2 1
Printed in the United States of America
on acid-free paper

For Nic and Dan

CONTENTS

ACKNOWLEDGMENTS

I find the process of creating economic ideas and writing papers to often be a painful process; it can take years and it can be alternately depressing and exhilarating. During this process, it is a comfort to have a coauthor. I owe a great debt to my coauthors, people I worked with on most of the papers in this volume. They are a very talented, hard-working collection of individuals, without whom I would never have undertaken or finished these papers. It was a pleasure working with them and they taught me a lot. Thank you.

How coauthored papers come about reveals some of the creative process that goes into producing ideas. In each case, a common language needs to be developed in order to communicate, describe, and understand new ideas. This can take many years or a few weeks. In each case, there has to be some prolonged contact so that ideas can simmer.

I met these coauthors in various ways, which shows how serendipity can play a role in research. For example, I met Frank Schmid in a bar in Vienna, where we first talked about corporate finance and banking. And Frank later visited the University of Pennsylvania so we could talk more. We ended up writing several papers together, one of which is in this volume. Don Mullineaux was the director of research at the Federal Reserve Bank of Philadelphia, where I worked first after leaving graduate school. Don gave me time to finish my thesis and also allowed me to spend bank time to repeatedly visit the archives of the New York City Clearing House Association. He was interested in financial history also, and we wrote a paper together. Lixin Huang and Ping He were two of my very best PhD students at the University of Pennsylvania. Papers with Lixin and Ping emerged after many long conversations in my office and could not have been written without their skills and hard work.

I was surprised when I was invited to join the National Bureau of Economic Research (NBER) many years ago because I came from a "fresh water" school (interior universities, Carnegie-Mellon, Rochester, Chicago and Minnesota; the center of a particular brand of macroeconomics) not a "salt water" school (universities on the coast, notably Harvard and MIT, associated with another brand of macroeconomics). At the time the NBER was dominated by salt water schools. In those days, you were assigned to a small group to work together presenting early ideas to one another during several weeks in the summer at Harvard. This went on for some years. My group of ten included Ben Bernanke and Joe Stiglitz.

And, it also included Jim Kahn and Charlie Calomiris. I wrote papers with Jim and Charlie. The one with Jim is in this volume.[1]

Rich Rosen I met when I visited the Board of Governors of the Federal Reserve System for a week, where Rich worked at the time. Discussions there led to several papers, including the one included in this volume. Guillermo Ordoñez was a junior professor at Yale when I joined Yale. He sat in on my PhD course, taught jointly with Tri Vi Dang (who I later wrote papers with) and so we developed a common language and wrote a series of papers, one of which is in this volume. Finally, I tend to talk a lot with people in offices near my own and this often leads to joint work. George Pennacchi was my neighbor when we first joined the Wharton School in 1984. We were trained very differently, but eventually we formed a language and wrote several papers, two of which are in this volume. Similarly, Nick Souleles was in the office across from mine at Wharton. We talked over a period of years and eventually wrote the paper in this volume.

The last essay in this volume was originally published in a book honoring Kalyan Sanyal. Kalyan was in my PhD class at the University of Rochester, and in a four-person study group, including me, that functioned throughout our graduate program years. I am quite convinced that I would not have made it through graduate school were it not for Kalyan teaching me economics. Kalyan was the star of our class and a person who cared deeply about using economics to have a positive impact on the world. He returned to India after graduate school. Unfortunately, he died relatively young.

Talk and collaboration are essential to the process of creating new ideas, either formally or informally, however it comes about. So, I am also very indebted to those with whom I learned from, and perhaps wrote papers with, but those papers are not in this volume. These include Tri Vi Dang, Bengt Holmström, Andrew Metrick, as well as many talented colleagues and graduate students at Wharton and Yale. Tri Vi Dang was visiting Yale during my first year at Yale. As often happens, by coincidence we started talking and exchanging ideas. We jointly taught a PhD course. This eventually led to joint work with Bengt Holmström and also Guillermo Ordoñez. Andrew Metrick had been my colleague at Wharton, but his office was on a different hall so we never had prolonged discussions during that time. We didn't do any joint work until we both joined Yale and had offices next to each other. Andrew is an energetic, smart guy, with great organizational skills, from which I have also benefited. I first met Bengt Holmström many years ago in his office at MIT where we had a lengthy discussion about the theory of the firm. I remember this very well but Bengt doesn't remember this at all, so maybe it didn't happen. I can't explain this disjuncture. We started talking again at the Jackson Hole Conference of the Kansas City Federal Reserve Bank in August 2008. And, we haven't stopped since. I owe a great debt to Bengt Holmström, who has influenced my thinking in countless ways and who is a joy to talk to.

My PhD thesis committee included Robert Barro and Robert King. Barro insisted that my thesis include an empirical chapter, a demand for which I am

1. The paper with Charlie is "The Origins of Banking Panics: Models, Facts, and Bank Regulation." It is included in a collection of Charlie's papers called *U.S. Bank Regulation in Historical Perspective* (Cambridge: Cambridge University Press, 2006).

forever grateful. Bob King was a constant support and got my thesis to the finish line. Finally, I should like to thank Stanley Engerman, my economic history professor at Rochester and member of my thesis committee. In graduate school I took U.S. economic history from Professor Engerman. I was the only student (a telling sign) and we met on Mondays from one to four in the afternoon. I wanted to study U.S. financial history but Stan insisted that we also cover important topics, like slavery and the Populist Movement. Those afternoons in Stan's office, surrounded by bookshelves where the books were two or three deep (but he always knew exactly where the particular book was), were a formative experience. From Stan I learned to think as a historian, to think of history itself as a process with a structure. Without that course I would never have gone to the archives of the New York City Clearing House Association or ventured into the decade-long process of collecting Free Banking Era data.

1

Introduction

> There are few subjects on which there is more loose theorizing than that of the origin and remedy of panics. These crises are commonly spoken of as accidental freaks of the markets, due to antecedent reckless speculation, controlled in their progress by the acts of men and banks who have lost their senses, but quite easily prevented, and as easily cured when they happen. These are the notions of surface observers.
>
> — Henry Clews, *Fifty Years on Wall Street* (1908)

You would not be reading this sentence were it not for the financial crisis of 2007–2008. Sadly, it is the reality of that event that perhaps makes this book relevant. This book collects many of the research papers on banks, banking, and financial crises which I worked on over the past 30 or so years, papers which gave me the framework for understanding the financial crisis of 2007–2008. By collecting these papers in one place I hope to convince the reader of the necessity of a historical vantage point for understanding the economics of banking and banking crises. The papers in this volume span almost 175 years of U.S. banking history, from pre–U.S. Civil War private bank notes issued during the U.S. Free Banking Era (1837–1863), followed by the U.S. National Banking Era (1863–1914) before there was a central bank, through loan sales, securitization, and the financial crisis of 2007–2008. During these 175 years, banking changed profoundly and yet did not change in fundamental ways. The forms of money changed, with associated changes in the information structure and infrastructure of the economy. Bank debt evolved as an instrument for storing value, smoothing consumption, and for transactions, but its fundamental nature did not change. In all its forms, it is vulnerable to bank runs, without government intervention. That did not change.

The message that short-term bank debt, in all its forms, is vulnerable to bank runs is delivered by financial history. The idea that financial crises are

fundamentally the same has been intuitively noted for over a century and a half, perhaps longer. For example, Ben Bernanke (2013): "The recent crisis echoed many aspects of the 1907 panic." *Wall Street Journal*, December 16, 1907 (p. 1): "In so many ways does the panic of 1907 resemble that of 1857." And on December 23, 1907: "[I]t is well worth while to compare the crisis of 1907 with that of 1873" (*WSJ*, p. 1). And so on. We also feel that financial crises are different, different from a recession or a stock market crash. There is no continuum of crises from mild to devastating. There are recessions, and other bad events, and then there are financial crises. It was repeated endlessly during the recent crisis that it was the "worst crisis since the Great Depression." And that is right. Crises are fundamentally different.

My PhD thesis of 1983, entitled "Banking Panics," looked at financial crises—bank runs—theoretically and empirically. The empirical work focused on the U.S. National Banking Era, 1863–1914, the period between the U.S. Civil War and the founding of the Federal Reserve System. Until the financial crisis of 2007–2008, there had not been a financial crisis in the United States since the Great Depression, yet I worked on this topic because I believed it was relevant for the modern world. The continuing recurrence of financial crises throughout the history of market economies strongly suggested to me that these events have a common cause, that there is something fundamental to be learned about the structure of market economies by studying financial crises. I persisted in research on these topics over my career as an economist because of the view that history is not a sequence of random events. There is some defining logic to market economies and to their histories. The past is relevant. Perhaps I have this view because I started in Marxist economics before I went to graduate school in (neoclassical) economics. To me the importance of history seems obvious. My PhD program required specialization in two fields; mine were macroeconomics and econometrics, but I added a third, economic history.

Financial history highlights the recurring episodes of financial crises in market economies. And, for hundreds of years, societies have pondered financial crises, banking panics. "After generations of theories, hypotheses and postulates, our economists today are still at odds over the causes of the familiar apparition, the Panic" (Collman 1931, p. 3). At some level, the basic problem has been understood for a long time. For example, Oscar Newfang (1908) writes:

> [The banker] promises to return deposits on demand, and then invests them in time obligations; so that no matter how good the paper which he has discounted, or how great an assurance he may have that the obligations will be met when due, he is not in a position to repay depositors, should they all desire their money immediately. (p. 728)

President Franklin Roosevelt also explained this in his first radio fireside chat, March 12, 1933, in the midst of the banking panics of the Great Depression:

> [L]et me state the simple fact that when you deposit money in a bank, the bank does not put the money into a safe deposit vault. It invests your money in many different forms of credit—in bonds, in commercial paper, in mortgages and in many other kinds of loans. . . . What, then, happened during the last few days of February and the first few days of March? Because of undermined confidence on the part of the public, there was a general rush by a large portion of our population to turn bank deposits into currency or gold—a rush so great that the soundest banks couldn't get enough currency to meet the demand. The reason for this was that on the spur of the moment it was, of course, impossible to sell perfectly sound assets of a bank and convert them into cash, except at panic prices far below their real value.

Banks issue short-term debt so that it can be a flexible store of value, depositors can write checks or withdraw any time. But, the assets of banks are longer term and cannot be readily liquidated if need be. This is a basic point of Douglas Diamond and Philip Dybvig (1983). But why would depositors all want their money at the same time? "[I]f there is the slightest doubt in [the depositor's] mind that the bank will meet its obligations *on demand*, he withdraws his balance" (Newfang 1908, p. 728). President Roosevelt attributes the runs to "undermined confidence." Still, this is not an explanation. What do "slightest doubt" and "undermined confidence" mean?

Explaining a financial crisis requires explaining why there is a sudden collapse of the financial system. The collapse of the financial system is a systemic event. Stock market crashes are not financial crises. The U.S. Savings & Loan crisis in the 1980s never threatened the entire U.S. financial system, although it was expensive to clean up. These are not systemic events. What does "systemic" mean? With respect to the recent financial crisis, Federal Reserve Chairman Ben Bernanke, in his Financial Crisis Inquiry Commission testimony, noted that of the "13 . . . most important financial institutions in the United States, 12 were at the risk of failure within a period of a week or two" (Bernanke 2010). The financial *system* was going down. This same point has been made about every panic. For example:

> At the present moment [during the U.S. Panic of 1837], all the Banks in the United States are bankrupt; and, not only they, but all the Insurance Companies, all the Railroad Companies, all the Canal Companies, all the City Governments, all the County Governments, all the State Governments, the General Government, and a great number of people. This is literally true. The only legal tender is gold and silver. Whoever cannot pay, on demand, in the authorized coin of the country, a debt actually due, is, in point of

> fact, *bankrupt*: although he may be at the very moment in possession of immense wealth, and although, on the winding up of his affairs, he may be shown to be worth millions.
>
> — (GOUGE 1837, p. 5; italics in original)

Without banks there is no money. In a crisis, cash is hoarded and bank checks are not acceptable. As Charles Fairchild, a member of the Monetary Commission and ex-Secretary of the Treasury put it, speaking of the U.S. Panic of 1893: "The thing that impressed me was the entire disappearance of all forms of money everywhere" (U.S. House Hearings 1897–98, p. 155). This was called a "currency famine" prior to the Federal Reserve's existence (Warner 1895). Following the collapse of Lehman Brothers there was also a currency famine.

Why do these crises occur? My PhD thesis of 1983 consisted of three papers, all published in this volume (though one is basically a new paper, written with one of my former PhD students, Lixin Huang). The three papers are discussed individually later. The basic point of my thesis was that a financial crisis—a bank run—is an *information event* which affects short-term bank debt. Holders of bank debt (depositors, for example) observe bad news about the future of the macroeconomy and become concerned that their bank might become bankrupt. Depositors know that most banks will be fine, but some will become insolvent, and their bank might be in trouble. All depositors reason this way and so they all run on their banks to withdraw cash. The depositors rationally react to unexpected news. Since my thesis, this basic story of crises has become much more refined.

The notion of bad macroeconomic news arriving triggering a crisis informed the empirical work. In the empirical work on the National Banking Era in my PhD thesis I determined what this news actually was, and I showed that the unexpected news had to exceed a threshold to trigger a panic; the news had to be bad enough. Not all banks are, in fact, bankrupt in a crisis, only a few (as I showed in the empirical chapter on banking panics during the U.S. National Banking Era). President Franklin Roosevelt also recognized this during the Great Depression: "Some of our bankers had shown themselves either incompetent or dishonest in their handling of people's funds. . . . This was, of course, not true in the vast majority of bankers" (First Fireside Radio Address, March 1933). Still there would be a bank run.

Following my thesis, the next 25 years of my research was largely concerned with further work on crises. My thinking has, of course, evolved, especially since 2007–2008, and I have a better understanding of financial crises than I did 30 years ago (I hope). In discussing the papers in this volume, I try to explain my thoughts at the time, but inevitably my current viewpoint projects backwards, putting the papers into alignment with my current thinking. Perhaps this is not so bad, but still I try to show the evolution of my thoughts at the time each paper

was written. While I indicate the year the paper was published, I do not discuss the papers in strict chronological order, but rather I try to give some overall logic by choosing an order that is based on the subject of each paper. The ordering is roughly historical.

This brings us to the first paper of this volume. The beginning question is why do banks and bank debt exist? What is "banking"? I tackled this question with George Pennacchi in "Financial Intermediaries and Liquidity Creation" (1990), the first paper in this volume (chapter 2). I start with this paper because it explains why banks exist. In this paper, we argue that the output of banks is debt used for storing value and trading. "Trading" means exchanging some form of "money" for goods or services. In this exchange, the "money" must be accepted by the other party. It would be best if the money was accepted without controversy, without questions and disputes about its value. Otherwise transacting would be very difficult. If the value of the money is not mutually clear and unquestioned, then one party to the transaction can take advantage of the other party because he may secretly have better information—this is called adverse selection. Transactions would be difficult to undertake. This was exactly the problem that existed when banks issued their own private money in the Free Banking Era in the United States.

Pennacchi and I equated "liquidity" with the idea of being able to transact without fear of adverse selection, that is, without worrying about some smart guy picking you off. Bank debt is created for this purpose. This debt must be such that there is no question about its value so that it can be used efficiently for trade. In the paper, the debt created by banks is actually riskless, making the point quite clearly. "The central idea of the paper is that trading losses associated with information asymmetries can be mitigated by designing securities which split the cash flows of underlying assets. These new securities have the characteristic that they can be valued independently of the possible information known only by the informed [party to a transaction]" (p. 50). Banks exist to create debt that is used for transactions.

The bank creates debt for trading purposes by contractually giving the debt holders the first rights to the bank's cash flows from the bank's loan portfolio. In fact, as long as the debt holders do not ask for their cash, the bank need not have the cash on hand, as Newfang and Roosevelt noted. The important point of the paper is that creating debt as senior to equity tranches (cuts) the information as well. Equity holders will be paid last, so they are very concerned about getting any money for their investment, making any information about the bank's loans is important for them. But, for the debt holders most information is not important because they are paid first. Consequently, most information is of no consequence to the debt holders, and everyone knows this, so debt can be used to transact without disputes. Bank debt separates the uninformed participants in the market from the privately informed, allowing the uninformed to trade

without any concerns about being picked off. Bank debt makes it so that possible secret information that the informed have does not matter.

"Financial Intermediaries and Liquidity Creation" was motivated by the widespread use of "noise traders" in financial economics. "Noise traders" are a theoretical construct, referring to economic agents posited to solve certain fundamental problems in financial economics.[1] The problem was posed (and solved) by Sanford Grossman and Joseph Stiglitz (1980): How can prices of securities be "efficient"—that is reveal or contain information—if private information is costly to produce? For some traders to be willing to spend resources to produce, and trade on, information, there must be some way for them to recoup their costs. If they are the "smart money," who is the dumb money? The role of "noise traders" (as they came to be known later) is to show up in the market and lose money on average when they trade, thus reimbursing the informed traders for their information production costs. Noise traders became a ubiquitous feature of financial economics.

Pennacchi and I asked ourselves how these "noise traders" would think. It seemed clear that they would want to trade with a security which was immune to losing money to insiders. This problem, of transacting with better informed parties, had been repeatedly discussed in history because it has been a problem for much of human history. For example, when coins were used, there was the problem of "shaving" off part of the gold or silver coin and then presenting the coin as whole. Of course, the coins could be weighed to determine their value (producing information), but then the question arises of whether the scales are fair, and there would be disputes over that. I had already studied the Free Banking Era, a period of U.S. history when this was a very important problem. We allude to this in the opening paragraph of the paper when we mention small, unsophisticated traders—"the farmer, mechanic, and the laborer" as corresponding to "noise" traders. In U.S. banking history, this association was often made, for example, New York State Legislature, *Report on Banks and Insurance* (1829): "The loss by the insolvency of banks generally falls upon the farmer, the mechanic and the laborer, who are least acquainted with the conditions of banks" (p. 14).

When the noise traders trade with a security that is vulnerable to sophisticated traders having more information than they do, they lose money. Historically, it has been difficult to find a way to transact without large costs being imposed by the form of money. With private bank notes, there is the same problem as with coins. When the notes of a bank circulate some distance away from the bank, their value becomes questionable and they would trade at discounts determined in a secondary market. But, what should the discount be? And who determines

1. See James Dow and Gary B. Gorton, "Noise Traders," *The New Palgrave: A Dictionary of Economics*, edited by Steven N. Durlauf and Lawrence E. Blume (New York: Palgrave MacMillan, 2008).

the discount? Thus, an essential feature of banking is that private money should be created that does not have these problems.

The equation of "liquidity" with a security that is immune from others having private information seems like a natural definition of liquidity. Another notion of liquidity comes from Douglas Diamond and Philip Dybvig (1983) and a third is due to Bengt Holmström and Jean Tirole (1998, 2013). Diamond and Dybvig associate "liquidity" with consumption insurance; depositors share the risk of consumption timing, ensuring that at some future uncertain date the value will be available for consumption. This also seems like a natural definition of liquidity. In my paper with Pennacchi, there is trade so the agents need to obtain goods in exchange for "money." In Diamond and Dybvig there is no trade, but instead agents might want to withdraw from the bank in order to consume (in effect withdrawing goods). If the agents wrote checks instead, in order to buy goods, then they would want the checks to be immune to adverse selection. The two notions of liquidity seem complementary since storing value and then "spending" it later is how things actually work. Holmström and Tirole think of liquidity as pledgeable cash flows, assets with cash flows that are readily verifiable. Pledgeable assets provide insurance against possible bad events in which agents need "liquid" instruments. An example is firms holding large amounts of short-term debt (money market instruments) or firms that sign up with banks for credit lines. Why can't these agents just sell other assets if needed? In Diamond and Dybvig liquidating the long-term project is costly and is best avoided. In Holmström and Tirole the problem is pledgeability; aside from pledgeability markets are complete. Some assets have return streams that cannot be pledged to other agents because these return streams are noncontractible. A good example is human capital. I cannot contract to provide all my best ideas to someone else. Return streams that can be pledged are "liquid." This too seems like a natural definition.

In fact, the three notions of liquidity seem interrelated. A firm or household holds funds in a money market mutual fund or a bank checking account so that the money can be used easily and flexibly. The fund or bank buys assets or makes loans, respectively, which are based on pledgeable return streams such as short-term debt. Firms and households can write checks on their fund accounts. Firms and households do not have sufficient pledgeable return streams, so they are willing to hold funds in low-yielding saving devices, such as checking accounts. As Holmström and Tirole say, they use the terms pledgeable income, liquidity, and collateral interchangeably.

Pennacchi and I argued that the output of a bank is debt; that is the bank's product, debt that has the feature that it can be safely used in transactions. And there is a demand for this debt, even in the case where the debt does not pay interest and is not, in fact, always able to trade at par, as during the Free Banking Era, discussed below. If the output of banks is debt, then it is obvious that, other

considerations aside, the famous Modigliani-Miller (M&M) theorem is violated by banks (see Modigliani and Miller 1958, 1961, 1963). The core of this theorem is an irrelevance proposition that states conditions under which a firm's choice of a capital structure, what debt, equity, and other instruments it uses to finance itself, does not affect the firm's value. Franco Modigliani (1980) explains the M&M theorem as follows:

> [W]ith well-functioning markets (and neutral taxes) and rational investors, who can 'undo' the corporate financial structure by holding positive or negative amounts of debt, the market value of the firm—debt plus equity—depends *only* on the income stream generated by its assets. It follows, in particular, that the value of the firm should not be affected by the share of debt in its financial structure or by what will be done with the returns—paid out as dividends or reinvested (profitably). (p. xiii)

A world in which there is a demand for bank debt to be used as money is not a world in which there are "well-functioning markets" in the sense that Modigliani means. The world analyzed by Gorton and Pennacchi (and Diamond and Dybvig and Holmström and Tirole) is not one that has such markets. The most important way in which banks are special is that their debt is a product, so banks would like to issue a lot of debt. This is why Milton Friedman (1959) argued that free banking, a system in which banks print their own money, would not work; they would print too much money. I discuss this issue below.[2]

"Financial Intermediaries and Liquidity Creation" was not about financial crises. Crises are not mentioned. The link between this paper and financial crises was made later by Holmström (2009) in the context of the crisis of 2007–2008. Holmström pointed out that the use of all forms of short-term debt do not require credit due diligence when used for trade. "They are low-information markets where trading is based on trust because there is no time for detailed evaluations. . . . [These securities] are not information sensitive" (p. 266). And Holmström pointed out that in our original paper, the bank debt was riskless and so there was literally no information that could be produced to benefit a sophisticated trader. But, banks cannot literally produce riskless debt; the debt is risky, in fact, potentially very risky. A macroeconomic news event may result in a financial crisis. The idea that a crisis is a situation in which bank short-term debt that is information-insensitive becomes information-sensitive developed later out of these observations, following Bengt Holmström (2009, 2012); see Dang, Gorton, and Holmström (2013).

2. Andrew Winton and I explore the implications of this for bank capital in "Liquidity Provision, Bank Capital, and the Macroeconomy," unpublished. In the paper with Winton the problem is that agents need debt for trading and using equity to trade is costly because of adverse selection. In that paper, debt is always riskless.

Historically it took a long time for banks to be able to produce debt that would be accepted without fear of adverse selection. If there is a fear of adverse selection, then bank money is not accepted at par in a transaction, that is a ten-dollar check is not accepted for ten dollars of goods. As I said above, a leading example of this is the period before the U.S. Civil War in which banks issued their own private currencies, the Free Banking Era of 1837–1863.[3] Studying this period was the basis for my thinking about bank money needing to trade at par, without being questioned. During the Free Banking Era, banks could not create debt that would be unquestioned in trade except when it circulated very close to the issuing bank. Free bank notes traded at discounts from par when they circulated away from the issuing bank. How did this system work?

An important banking system, often used in the past around the world, is a system in which banks print their own money.[4] Each bank issues its own currency. In the United States before the Civil War this was how banking worked. There were around 1,500 (genuine) currencies circulating during this period. The period is often described as chaos, for example, "The difficulties presented by the circulation of a chaos of currencies" (Pessen 1985, p. 145). How could it work? Why would the money be accepted? How could a new bank enter the money market? Was it chaos? I explored the U.S. Free Banking Era in two papers: "Reputation Formation in Early Bank Note Markets" (1996; chapter 3) and in "Pricing Free Bank Notes" (1999; chapter 4). These papers were based on an extensive set of bank note discounts found in a bank note reporter published monthly in Philadelphia prior to the U.S. Civil War. Finding the bank note reporter was hard and entering the data was also time-consuming. The project, in fact, took a decade.

The private bank notes of Philadelphia banks typically traded at par in Philadelphia, since they could easily be redeemed for cash if there was any question about these banks' solvency. The notes of more distant banks, banks in other states or cities (or Canada), traded in Philadelphia at discounts from par. So, for example, a ten-dollar bank note issued by a Philadelphia bank might only be worth $9.90 in Pittsburgh. In Philadelphia the discount on the notes of banks from the same distant location would usually be the same, but not always. These discounts were functions of the time it would take to return to the issuing bank to redeem the notes. But, this was not the only determinant. The riskiness of the issuing bank also mattered.

3. The period is called the "Free Banking Era" because 1837 was the year in which New York State passed a "free banking" law, which allowed for less restrictive entry into the banking business provided banks backed their monies with certain state bonds. Not all states adopted such laws. Nevertheless the period has come to be known as the Free Banking Era. Prior to 1837 banks also issued their own private currencies.

4. Schuler (1992) identified 60 national instances were multiple private currency has been issued.

Could banks in this era be "wildcat banks"? That is, could banks enter the business of banking, print money and use it to benefit themselves, ultimately absconding and leaving the holders of their money with worthless pieces of paper? It was the common view at the time, and since then, that wildcat banking characterized this period.

> At this time . . . any person who could raise a small amount of money was permitted to establish a bank, and allowed to issue notes for four times the sum raised. This being the case, many persons borrowed money merely long enough to exhibit it to the bank inspectors, and then borrowed money was returned, and the bank left without a dollar in its vaults, if, indeed, it has a vault on its premises. The result was that banks were started all over the Western States, and the country was flooded with worthless paper. These were known as the 'Wild Cat Banks.' . . . I began to think seriously of becoming a banker. I accordingly went a few days after to a printer, and he, wishing to get the job of printing, urged me to put out my notes. . . . My head being filled with the idea of the bank, I needed little persuasion to set the thing finally afloat. Before I left the printer the notes were partly in type, and I studying how I should keep the public from counterfeiting them. The next day, my Shinplasters were handed to me, the whole amount being twenty dollars; and, after being duly signed, were ready for circulation. . . . At first my notes did not take well; they were too new, and viewed with a suspicious eye. But through . . . a good deal of exertion on my part, my bills were soon in circulation.
>
> — WILLIAM WELLS BROWN (1853)[5]

This fictional characterization remained the dominant view for over a century. The revision of this view began with Rockoff (1974) and Rolnick and Weber (1983, 1984). Rolnick and Weber (1984) studied bank failures in states with Free Banking laws and those without free banking laws and showed that banks in Free Banking states failed when the value of the bonds backing their private monies declined precipitously. Rolnick and Weber showed that the backing collateral for money, the state bonds, was the driver of bank failures, not wildcats. Banks failed when the collateral declined in value. This may seem like an obvious point now, but it was not so obvious then considering that for the prior century or more the idea of wildcat banks was the dominant explanation for bank failures in the Free Banking Era.

5. *Clotel; or, The President's Daughter* is a novel by ex-slave William Wells Brown; it is a fictional account of two slave daughters of Thomas Jefferson, thought to be the first work of fiction in the United States by an African American.

William Wells Brown's character describes trying to get his new bank notes into circulation. In "Reputation Formation in Early Bank Note Markets," the question I explored was how the bank note discounts differed for new banks printing their own money compared to established banks. How does a new bank enter the market? The theoretical answer to this was provided by Douglas Diamond (1989) in an elegant paper about reputation formation. My paper, "Reputation Formation in Early Bank Note Markets" is essentially a test of Diamond's model. The basics of the model are worth briefly summarizing because it is very important in other settings as well, as I discuss below. In Diamond's model there are three kinds of potential borrowers in a loan market. There are good borrowers with a safe investment, there are bad borrowers with a bad investment with a low expected return but a high maximum return (a negative net present value project), and there is a group which can choose between the two projects. At the beginning, all the borrowers look the same and lenders cannot see what investment decisions the borrowers make. Thus, lenders cannot offer different interest rates to different borrower types. Having received a loan, at the end of each period, some borrowers will default. But not all borrowers who selected the bad project will default, so it will take time to learn each borrower's type. Over time the offered interest rate will be lowered for borrowers with a history of not defaulting; the lenders are able to discriminate between different types based on their default histories. The important point is that this learning creates an incentive for the borrowers with a choice of projects to choose the good project, not the bad project. Borrowers with a choice of investments have an increasing incentive to choose the good project because the cost of default increases over time—evolving so as to acquire a reputation, since the interest rate for nondefaulters is decreasing, relative the rate for those with a bad credit history.

A new bank opening in the Free Banking Era similarly has a choice of backing their money with safer assets or riskier assets (or holding a smaller amount of reserves).[6] A new bank upon opening would have to have its money accepted, even though no one had seen it before, as described above by William Wells Brown (1853). Imagine someone offers you a piece of paper that looks like money; it has $10 engraved on it with an engraving of, say, a railroad. You have never seen such "money" before. Why would you take this note in exchange for your goods? I showed that the monies of new banks had higher discounts than other banks at that location when the notes traded at a given distant location—Philadelphia. This created an incentive for holders of the new money to return and monitor that new bank by asking it to redeem its notes in cash. New banks

6. In states with Free Banking laws banks had to back their money with state bonds, but could choose the other assets. A bank could be riskier by choosing riskier state bonds and other riskier assets.

had to hold more cash because their money would return with a greater frequency than the established banks. This would happen for a while until it was determined whether the new bank was of the same risk as other banks at that location. The new bank had to establish a reputation and then it had an incentive to maintain it because its discount was lowered to equal that of other banks at the same location. In fact, the market was efficient in the sense that the discounts on the new banks that subsequently quickly failed were higher than the discounts on the notes of new banks that subsequently did not fail. Market participants could distinguish types fairly quickly.

What determined bank note discounts? Bank notes are perpetual debt obligations which offer the holder the right to demand cash in exchange for the note at any time. The right to demand cash at any time is a put option. The time it would take to return to the issuing bank from Philadelphia was the effective maturity of the option. The time it takes to get from Philadelphia to any other location can be calculated with pre–Civil War travelers' guides.[7] In "Pricing Free Bank Notes" I showed that the embedded put option—the right to go back and ask for cash—allows for the recovery of the implied volatility on the notes of banks at given distant locations. "Implied volatility" can be calculated once it is recognized that a free bank note can be priced with the Black-Scholes option pricing formula.[8] And, in fact, the implied volatility, a measure of bank risk, does move with other measures of risk, such as the type of banking system—free banking or not, and whether branch banking was allowed or not. Also, some states had insurance systems for bank notes. Further, technological change, such as the introduction of the railroad, occurred during the period and improved transportation. This caused the effective maturities to decline, and this was incorporated into note discounts and implied volatilities.

The private bank note system was efficient in the sense of financial economics; that is, information was reflected in the note discounts so in that sense the notes were priced correctly. But, it was very economically inefficient for transactions.[9] Trying to buy goods and services with free bank notes was hard due to disputes over the value of the money. This type of complaint was commonplace during the Free Banking Era. Here is a description of the problems from D. R. Whitney:

7. In Gorton (1989), I calculated these distances based on the type of transportation using Disturnell's *A Guide between Washington, Baltimore, Philadelphia, New York, Boston, etc. etc.* for various years.

8. See Black and Scholes (1973).

9. These two concepts of "efficiency" are not synonymous. "Economic efficiency" is a well-understood term and is related to the Fundamental Welfare Theorems of economics. "Market efficiency" means that in a financial market the security prices reflect all available information. See Dow and Gorton (1997).

> The business man of today knows little by experience of the inconvenience and loss suffered by the merchant of sixty years ago arising from the currency in which debts were then paid. Receiving payment in bank notes, he assorted them into two parcels, current and uncurrent [*sic*]. In the first he placed the notes issued by the solvent banks of his own city; in the other the bills of all other banks. Upon these latter there was a discount varying in amount according to the location and credit of the bank issuing them. How great the discount he could learn only by consulting his "Bank Note Reporter," or by inquiring at the nearest exchange office. He could neither deposit them nor use them in payment of his notes at a bank. The discount on them varied from one percent upwards, according to the distance the bills had to be sent for redemption and the financial standing of the bank by which they were issued.
>
> — (Quoted by KNOX 1903, p. 365)

There also was the widespread problem of counterfeits. Horace White:

> The heterogeneous state of the currency in the [eighteen] fifties can be best learned from the numerous bank note reporters and counterfeit detectors of that period. It was the aim of these publications to give early and correct information to enable the public to detect spurious and worthless bank notes, which were of various kinds, viz.: (1) ordinary counterfeits; (2) genuine notes altered from lower denominations to higher ones; (3) genuine notes of failed banks altered to the names of solvent banks; (4) genuine notes of solvent banks with a forged signature; (5) spurious notes, as of bank that had no existence; (6) spurious notes of good banks, as 20's of a bank that never issued 20's; (7) notes of close banks still in circulation.

The number of counterfeit and spurious notes was quite appalling. "Nicholas's Bank-Note Reporter" had 5,400 separate descriptions of counterfeit, altered, and spurious notes. (Quoted in *Sound Currency*, Vol. VI (1899), p. 148)

Perhaps the term "wildcat banks" should be thought of as referring to the plethora of problems that existed during this period, when money did not trade at par.

The bank note market can be (market) efficient in that the discounts are accurate, but this accuracy did not mean that transacting was easy. Quite the opposite. The legal history of the pre–Civil War Era is replete with disputes about bank notes. Because of shortages of gold and silver, contracts were often written in terms of payment to be made in "current bank notes." But, then because note discounts varied over time and space, the meaning of this obligation was not always clear. For example, in *Smith v. Goddard*, a case that came before the Supreme Court of Ohio in 1823 (1 Ohio 178; 1823 Ohio Lexis 33),

the court wrote that "In the ordinary course of business bank notes or coin at the election of the debtor were tendered and received without distinction or hesitation. . . . The parties to this contract, by the expressions, 'to be paid in current bank notes such as are passing', could not have intended bank notes of equal value to specie." The problem then was determining which bank notes are "current." Testimony in *Pierson v. Wallace*, before the Supreme Court of Arkansas in 1847 (7 Ark. 282; 1847 Ark. Lexis 10), illustrates the problem. Plaintiff "in order to establish the value of current bank notes introduced Wilson, as a witness who stated that current bank notes . . . were specie paying notes—such as were at par—that there were in circulation. . . . Alabama notes, which were at a discount of fifteen per cent. and Missouri notes which were at par or very nearly so." And so on.

Bank notes were suspicious because it was not known if one party knew more about the true value of the note than the other party. The discounts on notes were determined in secondary markets for the notes, where note brokers traded notes and sometimes took notes for redemption. Since note brokers, the informed traders noted in Grossman and Stiglitz (1980), had to produce information about the banks, they would cover these costs by trading with "the farmer, mechanic, and the laborer." If "the farmer, the mechanic and the laborer," were "noise traders," then the "informed traders" were the note brokers. Appleton (1831):

> This state of [circulating private bank notes] introduced a new branch of business and a new set of men, that of money brokers, whose business it was to exchange these currencies, one for the other, reserving to themselves a commission of about 1/4 of one per cent.
>
> The state of the currency became the subject of general complaint, the brokers were denounced, as the authors of mischief. (p. 11)

As suggested by Appleton, the noise traders realized that they could be taken advantage of. So, there were all kinds of disputes about the value of bank notes, making transacting hard. One way to see this is by looking at legal disputes. In *Egerton v. Buckner*, a case that came before the Supreme Court of Louisiana in 1843 (1843 La. Lexis 108; 4 Rob. 346), the court "found that the plaintiffs were [note] brokers and were able to sell the notes at 72 cents on the dollar. The evidence showed that notes they purchased to return to the defendants had cost them only 60 cents on the dollar." Note brokers, the informed traders, could apparently do very well—at the expense of the uninformed.

In "Financial Intermediation and Liquidity Creation," Pennacchi and I argued that there was a demand for bank debt because it had advantages in its use as money. The pre–Civil War system of private bank notes shows that there is a "convenience yield" associated with this bank debt. These notes did not pay

interest, but nevertheless they were used because they provided a service to the holders: they could be used as a means of payment. And this was the case despite the costs imposed by trying to transact with disputes about the discounts. This was recognized at the time:

> A bank note is a bill of exchange payable to the bearer at sight. It is a title deed to a certain amount of coin, at a certain place mentioned and described in the note, the possession of which coin may be had, whenever it is demanded. But, instead of demanding the coin, and carrying it about in a bag, I find it more expedient to carry the note in my pocket. In Boston, a Boston bank note passes in all commercial transactions the same as coin, because everybody knows that should the holder of the note happen to want the coin, he has only to step into State Street, present his note at the bank, and carry the coin off at his leisure. But, a Philadelphia bank note does not pass in Boston, in the same way. Few people in Boston want coin in Philadelphia; and nobody wants the trouble of going to Philadelphia to get the coin described in the note, and the additional trouble of bringing it to Boston.
>
> —(Hildreth 1840, p. 139)

The description of the private bank note market by Milton Friedman (1959), which I quote in the Introduction of "Pricing Free Bank Notes," that such a fiduciary currency could not work, was not the case. People did use private bank notes as money despite the difficulties.

Gradually, a new form of bank debt grew significantly prior to the Civil War: checking accounts, also called demand deposits. And, after the Civil War, private bank notes were taxed out of existence as part of the National Bank Acts.[10] This transition from bank notes to demand deposits took economists a long time to understand. Bray Hammond (1957), in his Pulitzer Prize-winning book *Banks and Politics in America*, wrote, "the importance of deposits was not realized by most American economists . . . till after 1900" (p. 80). Hammond goes on to discuss why the growing importance of demand deposits was overlooked. Later, I discuss another change in the money form that went unnoticed until the Financial Crisis of 2007–2008. The change from notes to deposits was a very important change in the form of bank money. In "The Development of Opacity in U.S. Banking" (2013; chapter 5), I trace this transformation of bank debt and the banking system. It involved a very important change in the information environment of banking. Efficient markets reveal information—information leakage.

10. Some argue that were it not for this tax, private bank notes would have survived. Of the roughly 60 or so private money systems in the world, none survived, suggesting that private bank notes were dominated by demand deposits.

The bank note discounts revealed information about bank risk. It is usually assumed that market efficiency is desirable. In fact, when it comes to bank money it would be economically efficient if markets did *not* reveal information, related to the point of my paper with Pennacchi, but more closely related to Dang, Gorton, and Holmström (2013). Then there would be no disputes about the value of the money and it would be easy to transact.

In order for bank money to trade at par, information leakage causing uncertain note discounts had to be eliminated. Otherwise bank checks would not be accepted at par. Information might also be revealed by a bank's stock price. A decline in a bank's stock price might trigger a run on that bank. This is what the bankers themselves worried about when checks replaced notes. Here, there were two sources of information leakage. The banking system endogenously transformed to eliminate these leakages. First, with checks there were no longer any note discounts revealing bank risk. No secondary market could develop because checks were the joint liability of the person writing the check and the bank. There were not enough of an individual's checks to make it profitable for a secondary market to develop. Second, the markets for bank stock, active before the U.S. Civil War, endogenously became very illiquid, with little trade, a minimal information leakage.

The *endogenous* closing of informative bank note and informative bank stock markets allowed demand deposits to trade at par, at first only in cities, but eventually nationally. This development of opacity is an important feature of bank debt and banks.[11] There were no markets to trade bank liabilities; there was no incentive to produce information about banks. Bank notes could return to the issuing bank via note brokers who bought them in secondary note markets. But, the secondary market for bank demand deposits was internalized by private bank clearinghouses, where checks were cleared. Bank checks inherently involve clearing, the movement of checks from receiving banks to the banks where the obligations were redeemed. The easiest way to do this was for all the banks to meet at a central location and net each other's checks (i.e., to "clear" the checks). In other words, at the central location banks met sequentially and pairwise, aggregated all the claims on each other bank and then transferred the difference in cash to each other bank. Clearinghouses would become the bank examiners and monitors.

Once deposit insurance was adopted, bank stock could trade (more frequently). The information revealed in stock prices would not affect demand deposits and they would not trigger bank runs. Later, with the development of "shadow banking," bank money changed again and the issue of information leakage would again arise. The new forms of bank money were sale and repurchase

11. See Dang et al. (2014).

agreements (repo) and asset-backed commercial paper (ABCP) (short-term debt backed by portfolios of securitized loans in the form of bonds, called asset-backed securities (ABS)). As I discuss below, securitization was essential for these forms of money to function because opacity of the ABS allowed repo and ABCP to function as money.

Whether bank money was private bank notes or demand deposits (or, indeed, repo or ABCP), there were banking panics. Above, I described bank runs as information events. I first articulated this in my job market paper which I presented at various universities when I was looking for a job as an assistant professor, "Bank Suspension of Convertibility" (1985; chapter 6). In this paper, depositors receive a noisy bad macroeconomic signal about bank assets and since they do not know which banks are exposed to the negative shock, they withdraw from all banks. That is, without bank-specific information, the depositors become concerned about all banks when bad public news arrives. But, not all banks are actually insolvent. To keep from liquidating the banking system, banks "suspend convertibility"; they refuse to honor their debt contracts by exchanging cash for checks or notes. Banks simply refused to give depositors their cash. And, although this was illegal historically, the laws were never enforced. It was recognized that in a financial crisis, to save the banking system, debt contracts should not be honored. I explain the history of this in my book *Misunderstanding Financial Crises* (2012).

In "Bank Suspension of Convertibility" (1985), I argued that suspension was in the interests of banks and depositors. The problem was that depositors did not know which banks were insolvent even if there were only a very small number of insolvent banks. A small risk of losing your life savings could trigger runs. In "Bank Suspension of Convertibility," I described suspension as part of an implicit contract between the banks and the depositors. Neither the solvent banks nor depositors want to force sound banks into bankruptcy by liquidating their longer term loans. This is why suspension was often welcomed. "The suspension of Specie payments had the effect, presently after it took place, to calm, in some degree, the agitation of the public mind" (Hildreth 1840, p. 99, speaking of the Panic of 1837). Upon suspension there is investigation of the conditions of the banks to determine which banks are solvent and which are not solvent.

"Bank Suspension of Convertibility" left many, many questions unanswered. I said nothing about why banks exist nor did I convincingly explain bank runs. The paper is too simple in that it considers a representative bank, so the question of why all banks suspended jointly is not posed or answered. Also, the all-or-nothing feature of a bank run, that is, depositors withdraw everything or not, is a by-product of the way I modeled depositors. The depositor's utility function in the final period is risk neutral. As a result, they go to a corner solution: either they withdraw all their money from their banks or nothing. This is not a satisfactory or convincing story of bank panics.

In the conclusion of the suspension paper, I say that panics are an "information event" and that is the idea that I took to the data. Are panics information events, runs triggered by bad macroeconomic news? If so, what exactly is the news? This is an important question for understanding crises. Is the run caused by news about fundamentals, or is the run triggered by extraneous factors and then harming the economy? In "Bank Suspension of Convertibility," I argued that it was the former. The empirical work in my thesis, "Banking Panics and Business Cycles" (1988; chapter 7) addressed these questions. I focused on the National Banking Era in the United States, a period that has important advantages for research. It lasted from 1863 to 1914 and included five panics. While there were state chartered banks, the national banks, which included all the largest banks, were regulated at the federal level. So, to that extent, it was a homogeneous system. Also, there was no central bank, so there were no expectations of central bank action. This allowed the search for the news to have a chance of success. In other historical eras, this is very difficult. There are usually not enough panics over a fairly homogeneous period. And, the presence of a central bank affects depositors' expectations in ways that are hard to detect.

In order to undertake empirical work, a practical definition of a bank run is needed. In Charles Calomiris and Gary B. Gorton (1991), we proposed a definition. "A banking panic occurs when bank debt holders at all or many banks in the banking system suddenly demand that banks convert their debt claims into cash (at par) to such an extent that the banks suspend convertibility of their debt into cash or, in the case of the United States, act collectively to avoid suspension of convertibility by issuing clearinghouse loan certificates" (p. 96). As I discuss below, a clearinghouse loan certificate was a special kind of private money issued by the clearinghouse in times of panic. These certificates were the joint liabilities of the clearinghouse. This definition works for the U.S. National Banking Era because the issuance of clearinghouse loan certificates can be observed. The clearinghouse issues the certificates when widespread runs occur, and sometimes this act can calm depositors' fears. In "Banking Panics and Business Cycles," I use this definition. In other settings, defining a banking panic or a financial crisis is more complicated.

The empirical work aimed to uncover the information that arrived which would cause depositors to alter their expectations about the future and so run on their banks upon seeing the news. I wanted to find and measure the news that arrived, affecting expectations such that it caused the panic. Depositors believe their banks are fine most of the time and then suddenly change their beliefs such that they run en masse to withdraw their cash. Something happened to cause them to switch their beliefs from "no run" to "run." What happened? The empirical work was heroic since there were many, many econometric and measurement problems to face. The National Banking Era Comptroller of the Currency's *Call Reports* were not in machine-readable form, moreover much of the data had to

be hand-collected. Also, many variables had no corresponding data. There were only five banking panics during the U.S. National Banking Era to analyze. The difficulties illustrate the problems with doing research on financial crises.

Nevertheless I tried. I developed a small model of currency and checks, which gave me a first order condition (a decision rule) that involved a pricing kernel (measuring the relative benefits of consuming more today versus consuming in the future) for the currency-deposit ratio. Basically, when a depositor received news that a recession was coming, this was very important since all his savings were in the bank, a bank which might fail in the recession. The news meant that depositors might lose their life's savings just when marginal utility is high, in a recession. "Many . . . depositors had lost their life savings through bank failures in the panics of 1873 and 1884" (Noyes 1898, p. 191). Hence, the news triggered runs.

What could this news have been? There are many candidates; seasonal movements in short-term interest rates could spike sometimes. Also, panics were usually associated with the failure of a large firm, financial or nonfinancial. I looked at these possible explanations but I focused on the liabilities of failed nonfinancial businesses. My prior view was that this variable would be important because Arthur Burns and Wesley Mitchell (1946) had shown that this was a leading indicator of the business cycle. This variable was printed in newspapers, where it was also often discussed. I guessed that people in the economy would use this information as the basis for their expectations, changing their beliefs when there was an unexpected movement in this variable—news. This turned out to be right.

I showed that in the U.S. National Banking Era, panics happened *only* when the unexpected component of the leading indicator of a coming recession exceeded a threshold.[12] There were no instances where the threshold was exceeded without a panic. Moreover, the signal—a leading indicator of a coming recession, tended to arrive near business cycle peaks. Financial crises and business cycles are linked. And the view that crises are information events was confirmed. Importantly, few banks ultimately failed during and shortly after the crisis; the banking system was not insolvent. Nevertheless, without information about exactly which banks were the weakest, depositors ran on all banks.

The results allowed me to construct counterfactuals. What if after 1914, the year the Federal Reserve System actually came into existence in the United States, the Federal Reserve had not come into existence and there were bank runs whenever the news variable exceeded the threshold? I showed that there would have been a panic in the 1920s, June 1920, and in December 1929, the

12. I also studied banking panics during the National Bank Era jointly with Charles Calomiris (see Calomiris and Gorton 1991).

start of the Great Depression.[13] There was no panic in the 1920s and the panics in the Great Depression came later and were haphazard. The counterfactual is important because it shows how the presence of a central bank alters the timing or even the existence of panics. This is one reason why financial crises in the modern era can seem so different from historical panics. Although about 65 percent of the 147 financial crises since 1970 involved runs, they often came late, as in the Great Depression.[14] And, in the other cases governments intervened with blanket guarantees or nationalization. In the 1920s the existence of the Federal Reserve System and its discount window alone prevented panics, which was the purpose of setting up the Federal Reserve. In particular, the Fed's discount window would be available at all times, would allow secret borrowing by banks, and would essentially be backed by the government. Banks did avail themselves of the discount window in the 1920s. But the Fed introduced "stigma" to keep the discount window borrowing to a minimum. At the start of the Great Depression, although discount window borrowing is not publicly observed, depositors perhaps believed that banks would go the discount window. But the banks did not go to the discount window. And when large banks began to fail well after the bad news had arrived in December 1929, depositors started to run (see Gorton and Metrick 2014).

The Great Depression counterfactual helps explain modern financial crises, since the experience of delayed bank runs during the Great Depression became widespread subsequently. In most financial crises there are bank runs, but like during the Great Depression they occur late in the crisis. And sometimes there is no bank run, usually because the government or central bank has taken an action such as offering a blanket guarantee or undertaking nationalization of the banking system. It seems that bank debt holders expect central bank or government action, so they wait, and only run if there is no action. Consequently, the definition of a banking crisis has to be expanded to accommodate such expectations in modern financial crises. Laeven and Valencia (2012) collected data on 147 financial crises between 1970 and 2011. They define an event as a crisis if two conditions are met. First, there are "significant signs of financial distress in the banking system (as indicated by significant bank runs, losses in the banking system, and/or bank liquidations)"; second, there are "significant banking policy intervention measures in response to significant losses in the banking system" (p. 4). In the latter case, they define six measures as significant interventions.

13. The data I used was the U.S. Comptroller's *Call Reports*, which were based on bank examinations five times a year. There were no bank examinations in October 1929, the date when the stock market crashed. December was the next examination date.

14. See "Systemic Banking Crises Database: An Update," Luc Laeven and Fabian Valencia (2012), IMF Working Paper #WP/12/163.

Debt holders' expectations make studying modern crises difficult. But, still the problem is bank runs, either actual runs or incipient runs. The financial crisis of 2007–2008 was not like the usual crises that have occurred during the era of central banking. Rather, those bank runs looked like nineteenth-century bank runs. The bank runs involved new forms of bank debt, sale and repurchase agreements, and asset-backed commercial paper.

Although there have been financial crises involving runs on other forms of bank money (bills of exchange, private bank notes), most of the experience is with runs on demand deposits. Demand deposits are special because checks must be cleared. Consequently, private bank clearinghouses arose. The process of clearing means banks would be exposed to the risk of other banks not being able to meet their obligations in the clearing process. Consequently, individual banks had incentives to monitor the other members. As a result, the clearinghouse introduced membership requirements, bank examinations, disclosure requirements and other rules, and became a quasi-central bank during crises. The opacity of banks due to the elimination of information-revealing markets meant that there would have to be nonmarket-based discipline. Information-revealing securities markets are often thought to create "market discipline," that is, the weaker firms or banks are revealed and must pay more to borrow, for example. But, bank checks relied on a lack of information, so the clearinghouse took the role of disciplining member banks. That is why there can be no discussion of demand deposits without a discussion of clearinghouses.

I began studying clearinghouses in the 1980s by exploring the archives of the New York City Clearing House Association. Two papers explain my findings: "Clearinghouses and the Origin of Central Banking in the U.S." (1985; chapter 8) and "The Joint Production of Confidence: Endogenous Regulation and Nineteenth Century Commercial Bank Clearinghouses" (written with Donald Mullineaux, 1987; chapter 9). These papers explain how the clearinghouse worked, especially during bank runs. Clearinghouses had to address the fundamental irony of privately produced bank money, namely, that the money was designed to be opaque and yet this very characteristic led depositors (or note holders or indeed holders of any short-term bank debt) to run en masse if bad news arrived. What could the clearinghouse do to "restore confidence" in bank money?

Facing runs, there was suspension of convertibility. Then the clearinghouse issued "clearinghouse loan certificates," effectively private money which was the joint liability of the member banks. Borrowers' identities were kept secret. Also, bank checks certified "Only Payable through the Clearinghouse" also operated as joint liabilities. Effectively, the member banks became a single institution, a large single diversified bank, meaning that debt holders did not need to worry about whether their individual bank was insolvent. The transformation of the

member banks into a single institution, issuing joint liabilities and consequently only revealing aggregate information, was truly remarkable.

During the suspension period, a new market opened to reveal the risk that the clearinghouse was insolvent. This was the market for cash in terms of certified checks. Newspapers reported the currency premia on certified checks, that is, the dollar value of certified checks that had to be paid for a dollar of cash. For example, a 3 percent currency premium meant that it took $1.03 dollars of checks to buy a dollar of currency. This is akin to the private bank note discounts, but now applied to the banking system (particularly since the large banks in New York City were effectively the banking system). In other words, an informative new market was created during crises, but one that revealed the risk of the entire banking system, not the risk of individual banks. When the currency premium reached zero, the crisis ended. See Gorton and Tallman (2014).

How did the institution of the clearinghouse work? In the third chapter of my PhD thesis, rewritten with Lixin Huang, "Bank Panics and the Endogeneity of Central Banking" (2006; chapter 10), we theoretically argue that when the dominant form of bank money is checks, private bank clearinghouses necessarily form and take on a central banking role in a banking system with many banks. And, importantly, during a panic the clearinghouse member banks join together to act as a single bank. This coalition of banks must be incentive-compatible which requires that banks monitor each other during normal times. Each bank knows that there is the possibility of a bank run in the future. Then, in order to keep themselves from being liquidated, the banks would have to act as one. Foreseeing this, the banks had incentives to mutually monitor in advance of the panic so that, as a group, they would be strong enough to survive and recreate confidence. The effect of forming a coalition when there is a run is informational. In response to the news shock causing the run, the coalition forms into a single bank portfolio, diversifying the risk that any individual bank is insolvent.

But, this also meant that clearinghouses could not prevent panics. In order to have incentives to mutually monitor, depositors had to monitor the banks periodically, that is, run on the banks to see if the coalition was, in fact strong. Since panics are costly, it would be best to avoid them altogether, which requires a central bank or deposit insurance.

We also showed that the industrial organization of the banking system is critical to the efficiency of dealing with banking panics. The most efficient banking system is one with a few large banks—ironically given the to-do about "too-big-to-fail." When there are many small banks, the clearinghouse system can approximate this. Over time, the industrial organization of a banking system can change, with new forms of financial institutions and new forms of money. These new institutions may not be regulated institutions.

Clearinghouses no longer deal with financial crises; governments and central banks have taken over this role. And, governments have taken over the role of

examining and regulating banks to discipline them. The idea is that government institutions, with credible discount windows that are always available and with strict bank examinations, can do what private clearinghouses cannot do, prevent banking panics. Government institutions can create confidence that the money will always be there. Clearly, things have not always worked out this way and there are financial crises, with the government or central bank responding with bank bailouts.

Why do governments or central banks bailout private banks occur during crises? In the recent crisis several large firms were bailed out, notably Bear Stearns and AIG. These bailouts of banks were not popular and led, in part, to the anti-banker backlash. In a crisis, the banking system is insolvent in the sense that no bank can honor its short-term debt. The question is whether the government or central bank should simply let the system be liquidated ("resolved" is the current euphemism). Bailouts in one form or another are inevitably the response of governments and central banks to crisis. No country has ever (intentionally) liquidated its banking system in a crisis. Prior to the Federal Reserve System, private banks bailed out clearinghouse members that were in trouble (see Gorton and Tallman 2014). In other words, it is not just governments that bailout banks in a crises.[15] This is very important to note because it means that bailouts are not obviously mistakes of governments, creating "too-big-to-fail" problems.

If banks get into trouble, as in a crisis when they cannot honor all the demands for cash, other investors should enter the market and buy these banks. It is a buying opportunity. For example, Bear Stearns' stock price was $133.30 the year before it was purchased at $10.00 per share. And there were some such purchases in the recent crisis. JP Morgan bought Bear Stearns and Washington Mutual, but with assistance from the government. Bank of America absorbed Merrill Lynch and Wells Fargo absorbed Wachovia. Barclays might have purchased Lehman, but in the end did not. However, the assets of the banking system are simply too large for private agents to buy, even at rock bottom prices. One only needs to look at the list of firms that received money under the government's Troubled Assets Relief Program and under the Federal Reserve's Term Auction Facility, the Primary Dealers' Credit Facility and Term Securities Lending Facility, not to mention the guarantee of all money market funds by the U.S. Treasury.[16]

In a financial crisis the whole financial system is teetering on the brink. The basic problem is that when the entire banking system needs to be sold, most resources in the economy are tied up in longer term projects and so are not available. Then there will be too little cash in the market, so even if the prices of firms

15. I also discuss this in *Misunderstanding Financial Crises* (Oxford: Oxford University Press, 2012).

16. Grossman (2010, chapter 4) reviews the history of bank bailouts.

up for sale fall, it may still be too much for the private sector to absorb, as with the case of AIG, for example. This is the maturity mismatch problem emphasized by Diamond and Dybvig (1983). In their model the problem is that consumers may want to withdraw from the bank before the banks long-term investments have reached fruition. If everyone does this, then the bank does not have the money, as Newfang and Roosevelt explained above.

In "Liquidity, Efficiency, and Bank Bailouts" (2004; chapter 11), Lixin Huang and I studied the role of the central bank when there is a systemic problem with the banking system. The question we address is why governments or central banks should, in fact, bail out their banking systems when there is a crisis. The reason that banking systems are bailed out is because of their role in the real economy. In our paper, there is a realistic link between the real economy and the banks. Banks lend to firms. If their borrowers get into trouble, the banks may have an incentive to simply roll over the borrowers' loans, for example. Banks should renegotiate the loans or liquidate the borrowers, but doing that has a negative knock-on effect for the bank then their bank may get into trouble, so the bank may want to avoid this. Caballero, Hoshi, and Kashyap (2008) showed that this happened in Japan. The problem arises when all the banks in the economy essentially face this problem, that is, it is a systemic problem. Then there is a role for the government because the problem is too large for the private sector to cope with.

Basically, Huang and I show that the assets of the banking system can only be purchased by the central bank. It is simply not efficient for private agents to hold enough liquidity so that they are prepared to buy the assets of the banking system in a crisis. Think of it this way. In the recent crisis, about three trillion dollars of assets needed to be sold by financial institutions to meet their short-term debt obligations. The resulting fire sale prices were a buying opportunity for private agents. But, private agents did not have three trillion dollars readily available and so, in the end, the Federal Reserve System purchased two trillion dollars' worth and commercial banks and hedge funds purchased, roughly, a trillion.[17] Only the government can create "liquidity" in large amounts in a short time. The government can issue a security (a Treasury bill or money) and bailout the banking system and support this by taxation in the future. The government is special in this sense, a fact noted by many others (see, e.g., Holmström and Tirole 1998).

"Liquidity, Efficiency, and Bank Bailouts" does not explain systemic financial crises, but focuses on why there are bailouts if there is a crisis. In the paper the private sector could be prepared to bailout banks by holding enough short-term assets (cash). But, it is very costly for society to hold so much cash that it is in a position to buy the assets of the banking system, should there be a crisis.

17. These numbers are from He, Khang, and Krishnamurthy (2010).

Most resources should be invested in the real economy. But, then these assets are illiquid—the maturity mismatch problem that Diamond and Dybvig (1980) highlight. In their model it is not best for each person to simply hold the short-term goods. It is better to invest in the long-term project. This is why the idea of having banks only hold short-term loans is not good.

Aside from producing debt for trade, banks have some other activities on the asset-side of their balance sheets. Banks make loans and loans are not the same as corporate bonds that a firm might issue. There is a fair amount of empirical evidence suggesting that loans are different from bonds (e.g., Lummer 1989 and James 1987). How are loans different? There is a large literature on this, which includes addressing bank–borrower relationships that endure because of what the lenders learn over time about borrowers. There are also studies of the information content of bank loan covenants. But, a very important feature of bank loans is that they can be more easily renegotiated (see Gilson, Lang, and John 1990). Unlike bonds, which are sold to many different investors, a bank loan has a single (or lead) lender. A single lender allows for renegotiation with borrowers, one-on-one. A loan, on average, in the United States, is renegotiated every eight months, which amounts to four times during the length of the loan, on average (see Roberts 2012).

In "The Design of Bank Loan Contracts" (written with James Kahn, 2000; chapter 12), we examine loan pricing and design, given that loans are easier to renegotiate. The incentives of each side to the loan contract are not aligned necessarily when there is some observable, but nonverifiable news that arrives about the borrower's future returns. If the news is bad, the borrower might engage in adding more risk to the project to gamble for resurrection (a moral hazard). The lending bank might try to extract more from the borrower in the renegotiation if it is able to (another moral hazard). We study this two-sided moral hazard problem.

We show that many features of bank loan contracts emerge endogenously in the model, for example, bank covenants that are tighter than those in bond contracts, the seniority of the bank loan, an option for the bank to liquidate at any time (due to tight loan covenants), and most important, that the loan rate is not set to price risk but to minimize subsequent renegotiation costs. The loan interest rate is set to try to mitigate the two-sided moral hazard. As a result, renegotiated interest rates are not monotonic in borrower quality. After the loan has been signed, the news arrives. A borrower receiving good news will not add risk to the project, and the loan interest rate does not change. If the news is neither good nor bad, the bank may lower the interest rate to prevent risk from being added. And, if bad news arrives so that there is no way for the bank to prevent the borrower from adding risk, then the bank tries to help itself by extracting a higher interest rate.

If renegotiation is a desirable feature of loans, and if there are conflicts because of the two moral hazard problems, perhaps there is some other way to mitigate these conflicts. Maybe the bank should own equity in the borrower. Owning stock in the borrower has two potential effects. First, the bank may be able to prevent management from taking risks (to gamble in order to increase the value of the equity) in the face of bad news. The bank would have an incentive to prevent this to protect its loan. Second, and more generally, when managers are entrenched and make decisions in their own interests, and not the interests of the shareholders, outside blockholders of the firm's stock can mitigate this problem. Banks may also play this role and may be better than nonbank blockholders.

Many banks around the world hold large equity stakes in their borrowers. Ownership of equity stakes in firms by banks is prohibited in the United States but is quite common in countries with universal banks, like Germany. In "Universal Banking and the Performance of German Firms" (with Frank Schmid, 2000; chapter 13), we collected data to examine German universal banking, an alternative way to organize a financial system. Germany is not so stock market–centered. We empirically study how German banks affect the performance of German firms in which they hold equity stakes. Do they behave opportunistically, reducing firm value, or do they add value?

In the United States, corporate governance emanates from the one share–one vote system. Control rights through votes related to equity are more complicated in Germany. Nonbank equity blockholders' voting may be restricted. There is also codetermination for large firms; this legal requirement requires that supervisory boards of directors have one-third or one-half employee representation. In short, there is no direct link between cash flow rights and control rights.[18] Thus, while nonbank blockholding is widespread in Germany, and bank blockholding is not extensive, still banks have enormous power. We found that firm performance improves to the extent that banks have control rights. And, banks improve firm performance by more than nonbank blockholders.[19] This is some evidence that alternative corporate governance systems work coherently, but it does not compare the efficiency of financial systems, an interesting but difficult task.[20]

18. Frank Schmid and I looked at German codetermination in "Capital, Labor, and the Firm: A Study of German Codetermination," *Journal of the European Economic Association* 2, no. 5 (September 2004): 864–905.

19. In "Stock Market Efficiency and Economic Efficiency: Is There a Connection?" *Journal of Finance* 52, no. 3 (July 1997): 1087–130, James Dow and I discuss how a bank can replicate a stock market in allocating resources, but based on internal information flows rather than through an efficient stock market.

20. These issues are discussed by Allen and Gale (1995).

Banks are opaque and this is why they are regulated and examined (see Dang et al. 2014). With clearinghouses, banks examined members and knew a lot about each other, although this was kept confidential. In the United States since the Federal Reserve System came into existence in 1914, banks are regulated and examined by the Fed, the Comptroller of the Currency, the Federal Deposit Insurance Corporation (since 1934), and state regulatory authorities. With clearinghouses, the examining member banks knew about each other, although this information was not public. But, lack of public information about banks means that banks do not know much about each other. For example, a bank does not know about the borrowers of other banks or what other banks are charging on their loans. So, how do banks compete? I come to this question below, but first we take a detour into credit crunches.

Banks have underwriting criteria—lending standards—for determining whether a prospective borrower will be granted a loan. Could changing lending standards affect macroeconomic activity? One possible way is through credit crunches, events in which banks reduce the amount of lending, not because loans are not demanded but for some other reason. As Ben Bernanke and Cara Lown (1991) point out, "there . . . still is a notable lack of consensus about the importance of a credit crunch" (p. 205). In fact, it is not clear that there have been credit crunches, partly because it is hard to distinguish between bank loan supply and bank loan demand. Is it that no one wants to borrow (demand) or that banks do not want to lend (supply)? Bernanke and Lown looked at a credit crunch emanating from the banking sector due to a possible shortage of bank equity capital, so that banks potentially supplied fewer loans. They show the effects of lower bank capital are small. In fact, the literature overall on credit crunches has focused on changes in bank capital, with mixed, mostly weak, results.

Perhaps credit crunches are more subtle and not just driven by bank capital changes, but to the way in which banks compete. In "Bank Credit Cycles" (with Ping He, 2008; chapter 14), we take a different approach to credit crunches.[21] We do not focus on bank capital, but on the unique way in which banks compete, since banks are opaque even to each other, in particular, on how banks screen possible borrowers when in competition with other banks. Lending criteria amount to producing information about prospective borrowers, screening out bad risks and lending to good risks, by producing information about the potential borrowers. It is costly to produce information about prospective borrowers. A bank could hire better loan officers, let them take longer to study the borrower, provide more detailed information, and so on. Or, a bank could cut costs and just do a minimum amount of work. The problem for banks is that their competitors are also choosing the quality of the information to produce

21. Holmström and Tirole (1997) is yet another approach to thinking about credit crunches.

about borrowers. In the paper we write: "Banks produce private information about their borrowers, but they do not know how much information rival banks are producing. The information opaqueness affects competition for borrowers in that rivals can produce information with different precision. This causes the imperfect competition in banking to take a different form from other industries." If rivals spend more on screening prospective borrowers, then they will get the good borrowers, leaving rivals with a pool of potential borrowers that has been adversely selected; the remaining pool is of lower quality on average.

In an oligopolistic setting, banks may want to save money by not being very precise about screening, that is, have low lending standards. But then to avoid adverse selection rival banks must also have low lending standards. Ping He and I considered an infinitely repeated game between two banks. The banks tacitly collude to produce only a baseline amount of information, not producing more because that is costly; they have low lending standards to save on costs. The banks expect to have the same average profits and the same average losses on loans. But this does not always turn out to be the case because of randomness; some loans inevitably default. Suppose one bank has bad luck and has a lot of defaults in its loan portfolio; and that bank observes that its rival has done better, having fewer defaults. The bank may then believe that the rival has increased lending standards, leaving it with a lower quality pool of borrowers. If a bank believes that the rival is deviating from an equilibrium in which they have tacitly coordinated not to expend a lot on screening, then both banks (all large banks in the system) raise their lending standards, causing a credit crunch.

In other words, even if all banks are tacitly colluding to produce only the low cost amount of information, still it can happen that all banks switch to producing much more information. If banks switch to higher lending standards, then this results in some borrowers who were getting loans before not getting loans now, a credit crunch. This is an *endogenous* credit crunch that affects the amount of borrowing in the economy. And, it is due to how banks compete with each other because banks are opaque.

If that were the end of the paper, it might be viewed as a clever theory paper (because banks are competing and colluding on lending standards). But, is the channel for credit crunches that we identified important in reality? We went on to test the model, and that is the important part of "Bank Credit Cycles." Testing this type of model (an infinitely repeated game) is very difficult. Our tests are not like the usual tests based on structural models. Our approach is the same as the approach I took in "Banking Panics and Business Cycles," namely, to guess the information that banks use to form beliefs or expectations. What information do banks use to update their beliefs about rivals that can cause a credit crunch when the rival's results are better?

The only information about rivals that a U.S. bank, or anyone else, sees comes from the data the banks report to the bank regulators, the Comptroller of the

Currency's *Call Reports* in the United States. *Call Report* data on all banks is announced publicly on certain prespecified dates. We hypothesized that at those dates banks look at rivals' results using these data, form beliefs, and subsequently act on those beliefs. By looking at future bank loan performance, we could detect, on average, whether banks responded to what they saw in the data. We construct and examine indices of bank loan loss performance differences. If these differences increase, and banks switch from a low screening equilibrium to a high screening equilibrium, then subsequently banks should lend less and increase the quality of loans, resulting in lower loan losses and reduced profitability. We first looked at U.S. credit card lending and then at commercial and industrial loans. In both cases, the results were consistent with the model, much to our amazement.

If such endogenous credit crunches occur, then this is not a risk that can be hedged; it is a macroeconomic risk for the economy, and more so for small borrowers which have nowhere else to raise money. As such, this risk should be priced, that is, stock returns should reflect this risk, stock returns of *nonfinancial* firms. In an asset pricing context we form a mimicking portfolio for our parameterization of banks' credit histories and show that this is a priced factor. This factor is significant in explaining the stock returns of small nonfinancial firms (who mostly only borrow from banks) and for all sizes of banks (with traded stock). The way in which banks compete can affect the macroeconomy.

In the 1980s, U.S. banks became unprofitable due to competition from money market funds and junk bonds.[22] With competition from nonbanks, bank charter values fell.[23] Bank failures rose and a merger wave broke out.[24] "Charter value" refers to the intangible benefits from being a regulated bank, largely the monopoly profits from entry restrictions. What explains the rise in bank failures? The explanation for this that was put forward was that fixed-rate deposit insurance creates an incentive for banks to take on risk: moral hazard (see Keeley 1990). "Moral hazard" refers to the tendency of banks to take excessive risk because their deposits are insured, so the interest rate they pay on the deposits does not increase with risk. Part of the response of banks was to find new profit opportunities. For example, banks significantly increased commercial real estate lending. And banks increased their risk. Whether the increase in risk and bank

22. See "Money Market Funds and Finance Companies: Are They the Banks of the Future?" written with George Pennacchi, in *Structural Change in Banking*, edited by Michael Klausner and Lawrence White (Homewood, IL: Irwin Publishing, 1993).

23. Gorton and Winton (2003) survey the large literature related to moral hazard and the decline in charter value.

24. Merger waves are analyzed in a paper I wrote with Matthias Kahl and Richard Rosen, "Eat-or-be-Eaten: A Theory of Mergers and Merger Waves," *Journal of Finance* 64, no. 3 (June 2009): 1291–344.

failures was due to moral hazard or some other cause is a very important question in banking.

In "Corporate Control, Portfolio Choice, and the Decline of Banking" (joint with Richard Rosen, 1995; chapter 15), we argue that managerial entrenchment played a much more important role than moral hazard. "Entrenched" means that the managers own enough of the bank's stock to fend off outside shareholders' monitoring of their behavior, but not enough stock to want to maximize the value of the stock. Instead they engage in non-maximizing behavior. Entrenched managers maximize private benefits, returns that accrue to them but not to other stockholders. For example, David Yermack (2006) found that CEOs at companies that allow personal use of company planes underperform market benchmarks by more than 4 percent annually. Entrenched managers can earn private benefits of control. But, if the bank managers do not own enough stock to fend off outsiders, then they take lower amounts of risk and maximize profits. The same is true in cases where managers own a lot of stock. They too prefer to maximize the value of the stock and not take on inefficient risk. Thus, there is a trade-off between the private benefits of control and rewards of ownership which is complicated by being nonlinear. The relationship between ownership share and risk-taking is an inverse U-shaped function in theory. Because of this theorized nonlinearity, we tested this model with semi-parametric methods and found that managerial entrenchment rather than moral hazard was the explanation in this case.

There are important implications of this work. First, during the period from 1934 to the mid-1980s, there were few bank failures. This is because charter value was high. There was no moral hazard problem due to deposit insurance not being priced correctly. When charter value decreased starting in the mid-1980s, bank failures also started to increase. Our results mean that the corporate governance of banks is a particularly important issue when bank charter value is low. The problem is not deposit insurance.

Reduced bank profitability spurred financial innovation in banking during the 1990s. This innovation opened new markets for banks to sell their loans, rather than hold them passively on their balance sheets. In several papers I, together with co-authors, looked at different forms of financial innovation in banking, innovations that later would later grow to become very, very significant.

One innovation was loan sales. Banks began to sell commercial and industrial loans, loans made to firms, in large quantities in the 1990s. In "Banks and Loan Sales: Marketing Non-Marketable Assets" (1995; chapter 16), George Pennacchi and I analyzed this phenomenon. Loan sales are not supposed to happen according to existing banking theory. If banks can sell loans, then they have no incentive to screen or monitor borrowers so no one would buy the loans, goes the argument. With the changes in the 1980s, bank funding costs rose and it became profitable to sell loans. How is this incentive-compatible? In other words, why

would anyone buy a loan? We empirically explored two channels. First, the bank can retain a junior position in the loan, and second the bank can offer an implicit guarantee. We looked at a unique data set of 872 loan sales and found evidence consistent with both channels in operation.[25]

But, the contract used to sell loans does not include a provision requiring the bank to retain part of the loan, so the contract must be implicit. Loan sales, and securitization discussed below, are two more examples that are related to Douglas Diamond's reputation acquisition model, discussed above. As with free bank notes, if over time lenders can discriminate between bank types, it becomes more costly for a bank to misbehave. With loan sales, why don't banks sell their bad loans or securitize their bad loans? In the beginning, buyers of loans were very concerned about this and "required" that banks hold a junior piece of the loan, for example. But, over time as the lenders discriminated between banks based on their histories, there was a stronger incentive not to do this. Unfortunately, unlike in the Free Banking Era, there is no data available to test this for the modern examples. It is worth emphasizing that studying the Free Banking Era is not some arcane exercise. We can see Diamond's mechanism at work there, so it is reasonable to think the same mechanism could be working in other markets, although we do not have the data to test it.

Securitization, the process of issuing bonds backed by portfolios of bank loans, was another innovation. The resulting bonds are called "asset-backed securities" (ABS).[26] Securitization prior to the financial crisis was a very large and enormously important market. It started to become very significant in the 1990s and became a global market. Many countries adopted laws allowing for a legal entity that was tax neutral to facilitate securitization. Securitization involves setting up a special purpose vehicle (SPV) which issues bonds (asset-backed securities (ABS)) stratified by seniority and uses the proceeds to buy a loan portfolio. In "Special Purpose Vehicles and Securitization" (2006; chapter 17), Nicholas Souleles and I investigated securitization. This was another very labor-intensive data collecting and organizing project, which involved merging two Moody's data sets, without matching identifiers.

We wrote the paper as a primer on securitization, hoping that economists would become interested in the topic. They did, but only after the financial crisis, with a few exceptions. In the paper we ask: What is the source of value to selling loans off-balance sheet? What features of a special purpose vehicle are

25. Also, see "Are Loan Sales Really Off-Balance Sheet?" written with George Pennacchi, *Journal of Accounting, Auditing and Finance* 4, no. 2 (Spring 1989): 125–45.

26. I include in this general term residential mortgage-backed securities, commercial mortgage-backed securities, as well as securitizations of nonmortgage asset classes, like credit card receivables, auto loans, student loans, and so on.

important? And, do investors in the ABS expect that the sponsors (the firms originating the loans) will bail out their SPVs if the loan portfolio suffers problems? In other words, the last question asks whether there are implicit guarantees that are priced by investors.[27] These questions relate to the general issue of how securitization can be incentive-compatible. The institutional details are important here, in particular, the details concerning the fact that SPVs can effectively not go bankrupt. This means that the costs of issuing debt off-balance sheet would be lower, ceteris paribus, for risky firms.[28] Since that paper was written, securitization has grown to essentially include all large financial firms.

Using the data set of credit card securitizations, we showed empirically that riskier firms engage in more securitization; they are less likely to have income so losing the tax advantage of on-balance sheet debt is less important for them. This is less likely today as ABS issuers have acquired reputations, so the office balance sheet interest rates on the ABS are lower, I conjecture. More importantly, we showed evidence that investors implicitly price the implicit guarantee to rescue the SPV if need be. Investors in ABS ask for a risk premium for sponsors who are relatively more risky, because there is a higher chance that they will be insolvent when their SPV needs to be bailed out.

The changes in banking involving loan sales and securitization as responses to the decline of traditional banking were concomitant with a number of important global changes in finance. Global financial markets have become increasingly dominated by institutional investors—asset managers, pension funds, sovereign wealth funds, money market funds, and other banks—managing trillions of dollars. These investors and nonfinancial institutions developed a need for something like a checking account to save money and earn interest over a short period of time. It had to be set up for a short period of time because these entities might need access to the money. Furthermore, the mechanism for saving had to be safe. But, there are no government-insured checking accounts for this type of depositor. There is a low limit on what is covered by insurance at banks. This led to the growth of sale and repurchase agreements ("repo") and asset-backed commercial paper (ABCP). In repo the depositor receives collateral from the borrowing bank, which is returned when the repo matures, which is usually overnight or a few days. ABCP is short-term debt issued by a special purpose vehicle that holds a portfolio of asset-backed securities. Repo and ABCP came to be called the "shadow banking" system. The shadow banking system grew, that is a banking system in which large entities "deposited" money with dealer banks (the old investment banks) overnight for interest.

27. Indeed, during the financial crisis many banks did effectively guarantee their SPVs.

28. The tax advantage of the debt shield on interest costs, however, is lost when financing off-balance sheet.

These changes led to a demand for privately produced "safe debt" to use as collateral. "Safe debt" is what Pennacchi and I were talking about in "Financial Intermediaries and Liquidity Creation." In the 2000s, it was widely held that there was a shortage of collateral. Large amounts of U.S. government debt and agency bonds (issued by Fannie Mae and Freddie Mac) were held abroad, in China and the oil-producing countries, in particular.[29] For example, in 2001, the Bank for International Settlements noted this looming problem:

> With growth of collateral use so rapid, concern has been expressed that it could outstrip the growth of the effective supply of these preferred assets.... The increase in collateralized transactions has occurred while the supply of collateral with inherently low credit and liquidity risks has not kept pace. Securities markets continue to grow, but many major government bond markets are expanding only slowly or even contracting. The latter phenomenon was particularly evident in the United States in the second half of the 1990s. (p. 2)

Securitization grew as a response to the global demand for safe debt.

But, how could ABS be safe? In order for checks to circulate at par prior to deposit insurance, it was important that the markets for bank stock be illiquid, so the prices would not reveal information. How is information leakage stopped in shadow banking? ABS has several features that make it suitable for collateral. These features, reminiscent of the information environment created for demand deposits 150 years ago, make ABS opaque and hence the short-term debt that they back can trade at par. First, a securitization deal involves the issue of mostly AAA-rated debt, about 85 percent of the bonds issued that are linked to a specific portfolio are AAA. Junior to the AAA-rated debt are the lower rated bonds. But, importantly, there is no equity piece that is traded publicly; the equity is held by the originator (in various forms). Thus, there is no information revealed about the loans in the ABS backing the short-term debt. Second, the loan portfolios chosen for securitization transactions are homogeneous, that is, all credit card receivables or all prime mortgages, for example. Asset classes are never mixed. If asset classes were mixed, the correlation between the performances of these asset classes would be important and would create an incentive to produce information about the ABS. But, then the repo and asset-backed commercial paper could not function as money. Securitization is not about diversification; it is about the creation of collateral.

29. See Carol Bertaut, Laurie Pounder DeMarco, Steve Kamin, and Ralph Tyron (2011), "ABS Inflows to the United States and the Global Financial Crisis," Board of Governors of the Federal Reserve System, International Finance Discussion Papers, No. 1928.

Hence, on the supply side, securitization helped banks recover their profitability, while on the demand side, repo and ABCP grew, needing collateral. There was a growing supply of privately produced collateral that could be used as backing for these instruments. These issues are discussed further in Gorton and Metrick (2013).

The growth of repo and ABCP did not replace checks as the dominant form of bank money, since it involved a completely different clientele. Repo and ABCP became the way in which large institutions could obtain a safe way to store value, while earning interest. The growth of these markets was an outcome of the transformation of global financial markets into markets dominated by large players (asset managers, pension funds, sovereign wealth funds, money market funds, and other banks). This banking system, which again came to be called "shadow banking," was where the bank run occurred in the recent financial crisis. As I noted above, this crisis was particularly hard to understand because it was not observed by outsiders and because a conceptual basis for understanding it was lacking for those insiders who saw it. I confronted the problem of explaining the crisis to a wide nonacademic audience when I was asked to appear before the Financial Crisis Inquiry Commission (FCIC), a 10-person committee that had been appointed to investigate the causes of the 2007–2008 financial crisis. In "Q&A about the Crisis" (chapter 18), a short essay, I tried to offer an easily digestible explanation of the crisis.

The first question about the financial crisis that I addressed in "Q&A about the Crisis" is "What happened?" I start the answer this way: "This question, though the most basic and fundamental of all, seems very difficult for most people to answer." This seems to me to remain the case; most people cannot give a coherent answer. The dominant narrative is: "Greedy, immoral bankers created toxic assets which they sold to unsuspecting investors who relied on fraudulent credit ratings." This is not an explanation. For example, it does not explain why the United States did not experience a financial crisis between 1934 and 2007. In "Q&A about the Crisis," I suggested that an "explanation" of the financial crisis satisfy three criteria. The point of articulating these criteria was to raise the level of the conversation about the financial crisis. In this regard, I failed miserably.

But, in the economics profession affairs appear a bit different. The financial crisis of 2007–2008 alerted economists to the fact that such crises are not so rare (that they can be ignored), though perhaps infrequent. Yet, it is not only financial crises that are not rare. Credit booms are also not rare.[30] Financial crises are often preceded by credit booms; growth in credit prior to the crisis is the best predictor of the likelihood of crisis. So, a theory of crises should incorporate

30. In "Crises and Productivity in Good Booms and Bad Booms," Guillermo Ordoñez and I looked at 34 countries over 50 years and found that of the 1,700 years in the sample, 1,001 were spent in credit booms. Over 50 years, on average, a country spends 20.4 years in a boom.

credit booms. Macro models should be able to generate credit booms and crises. Also, current theories of crises are very unsatisfactory, not only because there is no credit boom, but because they posit that the crisis is the result of a "big shock." In "Collateral Crises" (2014; chapter 19), Guillermo Ordoñez and I present a theory of credit booms that might end in crisis, without resorting to a "big shock." This paper is based on the micro foundations of Holmström (2009, 2012) and Dang, Gorton, and Holmström (2013).

In the paper, we abstract from financial intermediaries and look at households lending directly to firms (banks do not stand between them). The loans must be collateralized. We had in mind a contract like repo backed by ABS as collateral. The quality of the collateral is not known, but it can be determined by producing information at a cost. It is not optimal for agents in the economy to produce information every period. They know the average quality of collateral. If information is not produced, then over time the perceived quality of all collateral starts to (rationally) look the same. As a result, more and more collateral is seen as being of (relatively high) average quality (while in reality there is still the same amounts of good and bad collateral), and more and more firms can borrow. Output and consumption rise in this credit boom. Everyone is happy.

In fact, the best outcome would be if no one ever produced information about the collateral. Then output and consumption would be at their highest levels. To approximate this, private agents will choose as collateral securities which are most likely to retain value and which are very hard to produce information about, securities like mortgage-backed securities, which are linked to land and relatively complex. Complex, privately produced, securities are best for collateral when there are not enough government securities available.

In this setting the effect of a *given sized shock* depends on the length of the credit boom. The longer the boom, the more bad collateral is being used, though which collateral is bad is not known. A bad news shock of a given size may have one of three effects. Nothing may happen, because the boom has not been protracted. Or, there could be a credit crunch in which firms scale back their borrowing to prevent information from being produced if the boom lasts longer, a credit crunch. Or, a longer boom ends in a crisis and information is produced. In both of these latter cases, output and consumption go down, more when information is produced.

The crisis in this setting is an information event. It is not optimal for everyone to produce information about collateral all the time, but only on occasion. Information may be produced in response to the bad news. One point the paper makes is that the crisis is not the result of an exogenous "big shock." In our paper shock size is fixed. And, then when it arrives, the size of the affect varies for a given shock size, as explained above.

The information event in "Collateral Crises" is the same as that in Holmström (2009, 2012) and Dang, Gorton, and Holmström (2013), since it is based on

ideas in those papers. What is added in "Collateral Crises" are the dynamics of learning or forgetting about collateral value. Over time, when information is not produced, there is a credit boom and the economy is increasingly fragile. Another important point is that systemic fragility is endogenous. Fragility builds up during the boom; there are not exogenous "tail events" that resemble big earthquakes.

The financial crisis has caused rethinking about financial intermediation theory and macroeconomics. In "Some Thoughts on the Recent Financial Crisis" (2014; chapter 20), I muse about these issues, some of which are clear (at least to me) and many of the issues need further research. The paper summarizes a lot of the themes of the other papers in this volume.

Why does any of this matter? Financial crises are devastatingly costly in human and economic terms. After each crisis, some "reforms" are made, similar to the Dodd-Frank Act (2010) in that they are believed to have solved the problem. "It had for many years been a cardinal doctrine, in American banking circles, that a panic like those of 1893 and 1873 would never again be witnessed in this country. The ground for this belief lay in the phenomenal increase of our economic strength" (Noyes 1909, p. 363, discussing the Panic of 1907). And then another one happens.

Another book entitled "The Maze of Banking" was published in 1863.[31] The author was "A Depositor." Here is the opening sentence of the book: "Study and research having inveigled us into the labyrinth of Banking and Banking Laws, the following Treatise shows how we have been 'in endless mazes of lost'." In the conclusion, A Depositor is not optimistic: "Panics, unfortunately, will come" (p. 63). One hundred and fifty years later, I would like to think we have made some progress. I hope the papers in this volume show that, perhaps pleasing A Depositor.

REFERENCES

Allen, Franklin, and Douglas Gale (1995), "A Welfare Comparison of Intermediaries in Financial Markets in Germany and the U.S.," *European Economic Review* 39, 179–209.

Appleton, Nathan (1831), *An Examination of the Banking System of Massachusetts* (Boston: Stimpson and Clapp).

Bank for International Settlements (2001), "Collateral in Wholesale Financial Markets: Recent Trends, Risk Management and Market Dynamics," Committee on the Global Financial System.

Bernanke, Ben (2010), "Causes of the Recent Financial and Economic Crisis," Statement by Ben S. Bernanke, Chairman, Board of Governors of the Federal Reserve

31. Published by W. P. Nimmo; Edinburgh and Glasgow. And by Simpkin, Marshall, and Co.; London.

System, before the Financial Crisis Inquiry Commission, Washington D.C. (September 2, 2010); see http://www.federalreserve.gov/newsevents/testimony/bernanke20100902a.htm.

Bernanke, Ben (2013), "The Crisis as a Classic Financial Panic," speech at the fourteenth Jacques Polak Annual Research Conference, Washington D.C., November 8, 2013. http://www.federalreserve.gov/newsevents/speech/bernanke20131108a.htm.

Bernanke, Ben, and Cara Lown (1991), "The Credit Crunch," *Brookings Papers on Economic Activity* 1991, no. 2, 205–47.

Bertaut, Carol, Laurie Pounder DeMarco, Steve Kamin, and Ralph Tyron (2011), "ABS Inflows to the United States and the Global Financial Crisis," Board of Governors of the Federal Reserve System, International Finance Discussion Papers, No. 1928.

Black, Fischer, and Myron Scholes (1973), "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy* 81, 637–54.

Brown, William Wells (1853), *Clotel; or the President's Daughter, A Narrative of Slave Life in the United States* (New York: Collier Books; 1970 reprint of original).

Caballero, Ricardo, Takeo Hoshi, and Anil Kashyap (2008), "Zombie Lending and Depressed Restructuring in Japan," *American Economic Review* 98, 1943–77.

Calomiris, Charles, and Gary B. Gorton (1991), "The Origins of Banking Panics: Models, Facts, and Bank Regulation," in *Financial Markets and Financial Crises*, edited by Glenn Hubbard, 93–163 (Chicago: University of Chicago Press).

Clews, Henry (1908), *Fifty Years on Wall Street* (Hoboken, New Jersey: John Wiley & Sons; 2006 reprint of original).

Collman, Charles (1931), *Our Mysterious Panics, 1830–1930* (New York: William Morrow & Co.).

Dang, Tri Vi, Gary B. Gorton, and Bengt Holmström (2013), "Ignorance, Debt and Financial Crises," working paper.

Dang, Tri Vi, Gary B. Gorton, Bengt Holmström, and Guillermo Ordoñez (2014), "Banks as Secret Keepers," working paper.

Diamond, Douglas (1989), "Reputation Acquisition in Debt Markets," *Journal of Political Economy* 97, 828–62.

Diamond, Douglas, and Philip Dybvig (1983), "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy* 91, no. 3, 401–19.

Dow, James, and Gary B. Gorton (1997), "Stock Market Efficiency and Economic Efficiency: Is There a Connection?" *Journal of Finance* 52, no. 3, 1087–1130.

Friedman, Milton (1959), *A Program for Monetary Stability* (New York: Fordham University Press).

Gilson, Stuart, Kose John, and Larry Lang (1990), "Troubled Debt Restructuring," *Journal of Financial Economics* 27, 315–53.

Gorton, Gary B. (1989), "Ante Bellum Transportation Indices," working paper.

Gorton, Gary B. (2012), *Misunderstanding Financial Crises* (New York: Oxford University Press).

Gorton, Gary B., and Andrew Metrick (2013), "Securitization," chapter 1 in the *Handbook of the Economics of Finance*, volume 2, Part A, edited by George Constantinides, Milton Harris, and René Stulz, 1–70 (Amsterdam, Holland: Elsevier).

Gorton, Gary B., and Andrew Metrick (2014), "The Federal Reserve and Financial Regulation: The First Hundred Years," *Journal of Economic Perspectives*, Winter.

Gorton, Gary B., and Guillermo Ordoñez (2013), "Crises and Productivity in Good Booms and Bad Booms," working paper.

Gorton, Gary B., and George Pennacchi (1989), "Are Loan Sales Really Off-Balance Sheet?" *Journal of Accounting, Auditing and Finance* 4, no. 2 (Spring), 125–45.

Gorton, Gary B., and Ellis Tallman (2014), "How Do Banking Panics End," working paper.

Gorton, Gary B., and Frank Schmid (2004), "Capital, Labor, and the Firm: A Study of German Codetermination," *Journal of the European Economic Association* 2(5) (September 2004).

Gorton, Gary B., and Andrew Winton (2003), "Financial Intermediation," in *The Handbook of the Economics of Finance: Corporate Finance*, edited by George Constantinides, Milton Harris, and Rene Stulz, 431–552 (Amsterdam, Holland: Elsevier).

Gouge, William (1837), *An Inquiry into the Expediency of Dispensing with Bank Agency and Bank Paper in the Fiscal Concerns of the United States* (Philadelphia: William Stavely).

Grossman, Richard (2010), *Unsettled Account: The Evolution of Banking in the Industrialized World Since 1800* (Princeton, NJ: Princeton University Press).

Grossman, Sanford, and Joseph Stiglitz (1980), "On the Impossibility of Informationally Efficient Markets," *American Economic Review* 70, 393–408.

Hammond, Bray (1957), *Banks and Politics in America from the Revolution to the Civil War* (Princeton, NJ: Princeton University Press).

He, Zhiguo, InGu Khang, and Arvind Krishnamurthy (2010), "Balance Sheet Adjustment in the 2008 Crisis," *IMF Economic Review* 58, 118–56.

Hildreth, Richard (1840), *Banks, Banking, and Paper Currencies* (Boston: Whipple & Damrell).

Holmström, Bengt (2009), "Comment on 'The Panic of 2007', by Gary B. Gorton," in *Maintaining Financial Stability in a Changing Financial System* (Federal Reserve Bank of Kansas City).

Holmström, Bengt (2012), "The Nature of Liquidity Provision: When Ignorance is Bliss," Presidential Address, Econometric Society, ASSA Meetings, Chicago.

Holmström, Bengt, and Jean Tirole (1997), "Financial Intermediation, Loanable Funds, and the Real Sector," *Quarterly Journal of Economics* 112, 663–91.

Holmström, Bengt, and Jean Tirole (1998), "Private and Public Supply of Liquidity," *Journal of Political Economy* 106, 1–40.

Holmström, Bengt and Jean Tirole (2013), *Inside and Outside Liquidity* (MIT Press).

House of Representatives, U.S., *Hearings and Arguments before the Committee on Banking and Currency of the House of Representatives*, Fifth-Fifth Congress, Second Session, 1897–98 (Washington, D.C.: Government Printing Office).

James, Christopher (1987), "Some Evidence on the Uniqueness of Bank Loans," *Journal of Financial Economics* 19, 217–35.

Knox, John J. (1903), *History of Banking in the United States* (New York: B. Rhodes and Company).

Laeven, Luc, and Fabián Valencia (2012), "Systemic Banking Crises Database: An Update," IMF working paper, WP/12/163.

Lummer, Scott (1989), "Further Evidence on the Lending Process and the Capital-Market Response to Bank Loan Agreements," *Journal of Financial Economics* 25, 99–122.

Modigliani, Franco (1980), "Introduction," in *The Collected Papers of Franco Modigliani*, edited by Andrew Abel, volume 3, xi–xix (Cambridge, MA: MIT Press).
Modigliani, Franco, and Merton Miller (1958), "The Cost of Capital, Corporate Finance and the Theory of Investment," *American Economic Review* 48, 261–97.
Modigliani, Franco, and Merton Miller (1961), "Dividend Policy, Growth, and the Valuation of Shares," *Journal of Business* 34, 411–33.
Modigliani, Franco, and Merton Miller (1963), "Corporate Income Taxes and the Cost of Capital: A Correction," *American Economic Review* 53, 433–43.
Newfang, Oscar (1908), "The Value of Time Deposits," *The Bankers' Magazine* 76, January to June 1908.
New York State Legislative Assembly (1829), *Report of the Committee on Banks and Insurance Companies*, made to the Assembly on February 13, 1829 (Albany, NY: Croswell & Van Benthuysen).
Noyes, Alexander (1898; revised 1909), *Forty Years of American Finance* (New York and London: G. P. Putnam's Sons).
Pessen, Edward (1985), *Jacksonian America: Society, Personality, and Politics* (Champagne-Urbana, IL: University of Illinois Press).
Roberts, Michael (2012), "The Role of Dynamic Renegotiation and Asymmetric Information in Financial Contracting," Wharton School, working paper.
Rockoff, Hugh (1974), "The Free Banking Era: A Reexamination," *Journal of Money, Credit and Banking* 6, 141–67.
Rolnick, Arthur, and Warren Weber (1983), "New Evidence on the Free Banking Era," *American Economic Review* 73, 1080–91.
Rolnick, Arthur, and Warren Weber (1984), "The Causes of Free Bank Failures," *Journal of Monetary Economics* 14, 267–91.
Schuler, Kurt (1992), "The World History of Free Banking," in *The Experience of Free Banking*, edited by Kevin Dowd, 4–47 (London: Routledge).
Warner, John DeWitt (1895), "The Currency Famine of 1893," *Sound Currency* 2, no. 6 (February 15), 1–20.
Yermack, David (2006), "Flights of Fancy: Corporate Jets, CEO Perquisites, and Inferior Shareholder Returns," *Journal of Financial Economics* 80, 211–42.

PART 1

Bank Debt

2

Financial Intermediaries and Liquidity Creation

GARY B. GORTON AND GEORGE PENNACCHI*

A widely held view is that the investor of modest means is at a disadvantage relative to large investors. This popular perception, dating from at least the early 19th century, has it that the small, unsophisticated investor—"the farmer, mechanic, and the laborer"—is least equipped to acquire information and is most often victimized by having to trade with better informed agents. U.S. history is repeatedly marked by incidents of real or imagined insider shenanigans and resulting popular initiatives against the "money trusts" and the "robber barons." This view is responsible for many institutions, e.g., the *SEC* antitrust legislation, and various forms of taxation. This argument has also influenced bank regulation where it has been used to justify government provision of deposit insurance as a matter of public policy.

The notion that informed agents can exploit uninformed agents has received some support from Kyle (1985) and Grinblatt and Ross (1985). They show that insiders can systematically benefit at the expense of uninformed traders when prices are not fully revealing. However, in these models the uninformed traders, called noise traders, are nonoptimizing agents; they simply trade and

* Both authors from Finance Department, The Wharton School, University of Pennsylvania. A previous version of this paper was entitled "Transactions Contracts." The comments and suggestions of Mark Flannery, Jeff Lacker, Chris James, Dick Jefferis, Bruce Smith, Chester Spatt, an anonymous referee, members of the University of Pennsylvania Macro Lunch Group, especially Randy Wright and Henning Bohn, and participants in the 1988 NBER Summer Institute, the 1988 Garn Institute Conference on Federal Deposit Insurance and the Structure of Financial Markets, the 1988 Winter Econometric Society Meetings, and the Federal Reserve Bank of Richmond were greatly appreciated. The first author thanks the NSF for financial support through #SES-8618130. Errors remain the authors'.

lose money. If informed agents can, somehow, systematically take advantage of uninformed agents, then it seems clear that the uninformed agents would be motivated to respond, possibly creating alternative mechanisms. In this essay we investigate whether financial institutions and security contracts will endogenously arise as a response to problems faced by uniformed investors with a need to transact. In particular, we ask whether there are a variety of solutions and whether government intervention might be a necessary feature of any of them.

We first consider an environment that is similar in spirit to the above traditional notion that investors might need to trade in markets where better informed agents are present. The uninformed agents in our model have uncertain consumption preferences but are optimizing agents. Like the previous research, we show that the informed agents may exploit the uninformed, even though here they are optimizing. However, this result holds only when certain contractual responses by the uninformed agents are precluded. We go on to consider how the uniformed agents would respond in order to protect themselves from losses to the insiders.

The central idea of the paper is that trading losses associated with information asymmetries can be mitigated by designing securities which split the cash flows of underlying assets. These new securities have the characteristic that they can be valued independently of the possible information known only by the informed. By using these securities for transactions purposes, the uninformed can protect themselves. While our focus is on trading contexts, Myers and Majluf (1984) have considered a related problem in corporate finance. When firm managers have inside information, the firm may face a lemons market in issuing new equity.[1] However, they show that, if a firm can issue default-free debt, then the firm does not have to pay a premium to outside investors. One conclusion of our paper, as discussed below, is that firms would be motivated to issue default-free debt even if there were no information asymmetries at the new issue date.

By focusing on information asymmetries within a trading context, we can develop a notion of a security's "liquidity." A liquid security has the characteristic that it can be traded by uninformed agents, without loss to insiders. We show how intermediation can create liquidity by splitting the cash flows of the underlying assets that they hold. By issuing debt and equity securities against their risky portfolios, intermediaries can attract informed agents to hold equity and uninformed agents to hold debt which they then use for trading purposes. The idea that intermediaries can alleviate the problem of trading against insiders provides a foundation for the demand for a medium of exchange such as money, which is often simply assumed in many monetary models (e.g., a cash-in-advance constraint).

1. Rock (1986) considers a similar problem.

Thus, we provide an argument for the existence of intermediation which is distinct from the previous literature. Recent research on the existence of intermediaries can be broadly divided into two literatures. One literature focuses on efficient lending arrangements when there exist information asymmetries between borrowers and lenders. Intermediaries are seen as the unique solution to such agency problems. Examples of research in this area include Diamond (1984) and Campbell and Kracaw (1980). Unlike this literature, which focuses solely on the asset side of intermediaries, our paper is similar to a second line of research which has investigated the properties of intermediaries' liabilities. In the seminal paper by Diamond and Dybvig (1983), banks provide liquidity by acting as risk-sharing arrangements to insure against depositors' random consumption needs. The intermediary contract prevents inefficient interruptions of production.

Like Diamond and Dybvig (1983), we are concerned with the idea that intermediaries provide liquidity. However, our notion of intermediaries as providers of liquidity differs in a number of important respects. As Jacklin (1987) and Cone (1983) have shown, a crucial assumption of Diamond and Dybvig (1983) is that agents cannot trade equity claims on physical assets. If a stock or equity market is open, this trading arrangement weakly dominates intermediation. Unlike Diamond and Dybvig, we do not arbitrarily rule out trading in a stock market. On the contrary, it is the presence of insiders in this market which motivates the formation of an intermediary. Second, our model differs in that the intermediaries here will explicitly issue debt and equity, serving as mechanisms that split cash flows. Finally, the existence of our intermediary does not rely on providing risk sharing or resolving inefficient interruptions of production. Our notion of liquidity as providing protection from insiders is fundamentally different.

Recent independent work by Jacklin (1988) is similar to ours in that, in the context of a Diamond and Dybvig-like model, he does not rule out trading in an equity market and shows that bank liabilities can prevent losses to informed insiders. However, the intermediary modeled by Jacklin does not issue debt and equity and is partly motivated on risk-sharing grounds. Our model differs in that intermediaries explicitly issue both debt and equity securities, thereby splitting the cash flows of their asset portfolio. Thus, in our setup, intermediaries explicitly create a new, liquid security. We also consider the feasibility of this intermediary contract by considering the conditions under which the intermediary can attract insiders to become equity holders. Thus, we justify the bank from first principles on grounds different from risk sharing.

Importantly, bank intermediation is not the unique solution for protecting uninformed agents. In our model, liquidity creation may be accomplished at the firm level without the need for bank intermediation. By issuing both equity and debt, firms can split the cash flows of their asset portfolios, thereby creating a

security (corporate debt) which is safer than their underlying assets. This debt can serve as the basis of a safe security that may be used by uninformed agents for transaction purposes.

A key point is that private transactions contracts may not be feasible under certain conditions. This might be viewed as a "market failure" from the perspective of the uninformed agents and could justify a role for government intervention. The government can intervene on their behalf in several ways. One way of protecting the uninformed agents is by insuring the deposits of the banking system through a tax-subsidy scheme. A system of government deposit insurance can achieve the same allocation as when private bank transactions contracts are feasible. Alternatively, if it is infeasible for corporations to issue sufficient amounts of riskless debt, government intervention in the form of a Treasury bill market can improve uninformed agents' welfare by providing additional riskless securities. This form of intervention is shown to parallel that of the provision of deposit insurance since, in both cases, the government's role is to create a risk-free asset.

The paper proceeds as follows. In Section 2.1 the model economy is detailed. In Section 2.2 a stock market allocation when all agents are fully informed is set out as a reference point. Section 2.3 considers the case of asymmetric information and shows how the informed agents can take advantage of uninformed agents by forming a coalition that trades in the stock market. Then, in Section 2.4 the private intermediary contract, when feasible, is shown to break the informed agents' coalition. When private contracts are infeasible, we show in Section 2.5 that government intervention by insuring bank deposits or creating a government debt market can be beneficial in protecting uninformed agents. Section 2.6 concludes.

2.1. THE MODEL ECONOMY

There are three dates in the model economy, $t = 0, 1, 2$, and a single consumption good. The following assumptions detail the model.

2.1.1. Preferences

There are three types of agents:

(i) Agents with known preferences at $t = 0$, who derive utility from consumption at date $t = 2$ given by $U = C_2$.

(ii) Agents with preferences that are unknown at date $t = 0$, but which are realized at date $t = 1$ to have utility from consumption at date $t = 1$ given by $U = C_1$ but no utility from consumption at $t = 2$. These agents are called "early" consumers.

(iii) Agents with preferences that are unknown at date $t = 0$, but which are realized at date $t = 1$ to have utility from consumption at date $t = 2$ given by $U = C_2$ but no utility from consumption at date 1. These agents are called "late" consumers.

Agents of types (ii) and (iii) will collectively be called "liquidity traders." Let N equal the number of liquidity traders, which is assumed to be large relative to the number of agents with known preferences. At $t = 1$ the proportion of liquidity traders with preferences for early consumption is realized. (The remaining fraction consists of late consumers.) The realized proportion of early consumers may be low, proportion w_l,which is expected to occur with prior probability q_l, or high, proportion w_h, expected to occur with prior probability q_h. It is assumed that $w_h > w_l$.

2.1.2. Endowments and Technology

At $t = 0$, all agents receive endowments of a capital good which when invested earns a return in the form of the consumption good at $t = 2$. Each liquidity trader is assumed to receive an endowment of one unit of the capital good, while type (i) agents with known preferences receive equal endowment shares of the capital good that total M units in aggregate. Capital is homogeneous, and each unit produces the same random return. Each capital unit produces either R_H units of the consumption good or R_L units of the consumption good at date $t = 2$, where $R_H > R_L > 0$. It is assumed that the probabilities at date $t = 0$ of each state occurring equal one half.

In addition to the capital good, all liquidity traders receive an endowment of e_1 units of the consumption good at $t = 1$, while type (i) consumers receive equal endowment shares of the consumption good at time $t = 2$ that total Me_2 units in aggregate. Each unit of the consumption good received by the liquidity traders at $t = 1$ can either be consumed at $t = 1$ or stored to yield a certain return of one unit of the consumption good at date $t = 2$.

2.1.3. Information Sets

At date $t = 1$, uncertainty about capital returns and liquidity traders' preferences is resolved. It is assumed that type (i) consumers have access to this information at date $t = 1$; i.e., they know whether the return on capital will be high or low and whether the proportion of early consumers in the economy is high or low. Thus, we will hereafter refer to the type (i) consumers as the "informed" traders.

While liquidity traders find out at $t = 1$ whether they are early or late consuming individuals we will consider the case where they are not directly informed about the aggregate proportion of early consumers and the realized return on

capital. In this case, information may or may not be revealed by the result of traders' actions at time $t = 1$. However, for purposes of comparison, we will first consider the "full-information" benchmark case where liquidity traders are assumed to directly receive information regarding the realized aggregate proportion of early consumers and the realized return on capital.

2.2. A STOCK MARKET WITH FULL INFORMATION

It is apparent that certain agents will desire to trade at $t = 1$. In particular, when some liquidity traders find that they are early consumers at $t = 1$, they will want to sell their entire endowment of the capital good for the consumption good at this time. In addition, other liquidity traders who discover that they are late consumers may want to sell their $t = 1$ endowment of the consumption good for the capital good if their expected return to holding capital is at least as good as their return to storing their consumption endowment. In general, the type (i) informed traders may desire to sell some of their capital good for the consumption good at time $t = 1$ in order to store it from $t = 1$ to $t = 2$. Whether informed traders want to sell capital will be an important issue when we consider the case of uninformed liquidity traders. However, it will become clear that ignoring the type (i) traders will not change the equilibrium for the full-information case.

Since each unit of capital invested at $t = 0$ is subject to the same source of risk (i.e., either all units produce a high return or all units produce a low return at $t = 2$), it will make no difference whether we think of agents individually investing their endowment of the capital good or giving it to firms who then issue to them shares reflecting a proportional claim to the capital's return at $t = 2$. Thus, a "stock market" is equivalent to individual investment of the capital good.

Let us then consider the stock market equilibrium in this full-information case. All agents' utility levels will be determined once we solve for the equilibrium price of the capital good in terms of the consumption good at date $t = 1$. We do this for the four possible states of nature realized at date $t = 1$; $\{i,j\}$, $i = h,l$, $j = H,L$, where i refers to a high or low proportion of early consumers, while j refers to a high or low return on the capital good. Let p_{ij} denote the date $t = 1$ value of one unit of the capital good in terms of units of the consumption good when state i,j occurs.

At $t = 1$ early consumers will wish to purchase the consumption good in exchange for their endowment of one unit of the capital good. Early consumers, in total, own Nw_i units of the capital good which they are willing to sell. The aggregate quantity of the endowment good demanded by the early consumers is Nw_ip_{ij}. Since the late consumers are the only agents from whom the early consumers can buy endowment of the consumption good, the late consumers will end up selling some or all of their endowment of the consumption good to the

early consumers. Let the amount of consumption good supplied by the late consumers be $S(p_{ij})$. If everything is supplied, then $S(p_{ij}) = Ne_1(1-w_i)$. Otherwise, some amount less than $Ne_1(1-w_i)$ will be supplied.

We now determine the price, p_{ij}, which clears the market at date $t=1$ in each state of the world $\{i,j\}$. Market clearing equates the demand for the consumption good with supply. Thus,

$$Nw_ip_{ij} \leq Ne_1(1-w_i). \tag{2.1}$$

There are two separate cases to consider, one where late consumers sell all of their consumption endowment (condition (2.1) holds with equality) and one where they sell only part, choosing to store some (condition (2.1) being a strict inequality).

When there is no storage in equilibrium, condition (2.1) becomes an equality. Solving for the price of the capital good, we have

$$\text{(No Storage)} \qquad p_{ij} = \frac{e_1(1-w_i)}{w_i}. \tag{2.2}$$

This case holds under the parametric restriction:

$$R_j > \frac{e_1(1-w_i)}{w_i}. \tag{2.3}$$

When storage occurs in equilibrium, late consumers must be just indifferent between buying and holding the capital good and storing the consumption good, i.e.,

$$\text{(Some Storage)} \qquad p_{ij} = R_j. \tag{2.4}$$

This case holds when the inequality sign in condition (2.3) is reversed.

Hereafter, we will make the assumption that condition (2.3) holds for $j = H$, so that, in equilibrium, no storage will occur for the states $\{h, H\}$ and $\{l, H\}$, where the return on capital is high. In addition, we will assume that condition (2.3) does not hold for $j = L$, so that, in equilibrium, some storage will occur for the states $\{h, L\}$ and $\{l, L\}$, where the return on capital is low. These assumptions can be summarized by the following condition:

$$R_H > \frac{e_1(1-w_l)}{w_l} > \frac{e_1(1-w_h)}{w_h} > R_L. \tag{2.5}$$

Condition (2.5) amounts to assuming a sufficiently high variance in asset returns relative to the variance in the proportion of early consumers. This assumption will lead to a more interesting problem when we consider the effects of asymmetric information.

Note that, for this full information case, type (i) consumers have no incentive to trade in the capital good at date $t = 1$. Whenever there is a high return on capital, the rate of return on capital exceeds that of storing endowment, so type (i) consumers will choose not to sell capital. When there is a low return on capital, the rate of return on capital just equals the return to storage, so that type (i) traders are indifferent to purchasing endowment.

Since type (i) agents do not trade, their expected utility (consumption) per unit of capital endowment at date $t = 0$ is

$$E[C_2] = e_2 + \overline{R}, \tag{2.6}$$

where $\overline{R} \equiv {}^1/_2(R_H + R_L)$.

The expected utility of liquidity traders can be computed from our previous results:

$$\begin{aligned} E[C_1 + C_2] = {} & \frac{q_h}{2}\left[w_h\left(e_1 + p_{hH}\right) + (1 - w_h)\left(R_H + \frac{e_1 R_H}{P_{hH}}\right)\right] \\ & + \frac{q_l}{2}\left[w_l\left(e_1 + p_{lL}\right) + (1 - w_l)\left(R_L + e_1\right)\right] \\ & + \frac{q_h}{2}\left[w_h\left(e_1 + p_{hL}\right) + (1 - w_h)\left(R_L + e_1\right)\right] \\ & + \frac{q_l}{2}\left[w_l\left(e_1 + p_{lH}\right) + (1 - w_l)\left(R_H + \frac{e_1 R_H}{p_{lH}}\right)\right] \\ = {} & e_1 + \overline{R}. \end{aligned} \tag{2.7}$$

In what follows, we will compare the expected utility of the different agent types under alternative information and trading settings to the expected utilities given by (2.6) and (2.7).

2.3. A STOCK MARKET WITH ASYMMETRIC INFORMATION

Now suppose the model is the same as that of the previous section except that only type (i) agents, the "informed traders," are assumed to have direct knowledge of the return on capital and the proportion of early consumers at $t = 1$. In this section we restrict liquidity traders to hold their wealth only in the form of stock. Given this assumption we ask whether the informed agents can collude at date $t = 1$ to exploit the liquidity traders. First, we summarize what will happen at $t = 1$. Then we define an equilibrium. Finally, we show the existence of insider trading in equilibrium.

The liquidity traders, early and late consumers, do not know what return capital goods will earn. Nor do they know the proportion of early consumers in the economy. At date $t = 1$, however, the decision of the early consumers is

straightforward. Regardless of possible information, they sell their capital goods for consumption goods. Late consumers must decide either to store their newly arrived endowments of the consumption good or to sell all or parts of these endowments for capital goods. This decision, made as a function of the market price, characterizes the behavior of the late consumers.

Informed agents know (as do all agents) that, in equilibrium, prices will reveal some or all information about the true state of the world. Consequently, they will need to coordinate their trading strategies (collude) in order to gain from their superior information. We assume that there is a sufficiently small number of informed agents such that they are able to form a trading coalition, if they individually so desire. Thus, at $t = 1$ the sequence of events is as follows. First, the informed agents communicate and choose an amount of capital goods that they will jointly supply in state $\{i,j\}$ knowing that uninformed agents will act competitively. We first solve this game between the informed agents. Then the equilibrium price is determined to clear the market between late consumers supplying endowment goods and early consumers, possibly together with informed agents, selling capital goods.

The amount supplied by the coalition in each state $\{i,j\}$ will be based on a strategy designed to make some states of nature indistinguishable from other states of nature when viewed by the uninformed agents. That is, the equilibrium prices in some states of nature will be the same as in other states of nature. In order for prices not to reveal the true states of nature in equilibrium, the optimal strategies of individual informed agents must be to supply no more capital goods than are supplied by the coalition acting on their collective behalf. The existence of the insider trading equilibrium will depend on showing that individual members of the informed agents' coalition have no incentive to deviate from the coalition strategy, by selling capital goods on their own unbeknownst to the coalition. In equilibrium it will be in the interest of each informed agent to be a member of the coalition and, once having committed capital for sale by the coalition, not to supply any additional capital. This is because, if any additional capital is supplied by individual informed agents (acting independently of the coalition), the equilibrium price will reveal the true state of the world. If this occurs, then no informed agent can benefit. We now briefly formalize this so that we can subsequently define an equilibrium.

Let $M_{ij} \leq M$ be the amount the coalition proposes to its members as the amount to be supplied in state $\{i,j\}$, with each member supplying an identical share. The coalition's strategy will be characterized by the amount of the capital good that the coalition supplies in state $\{i,j\}$, M_{ij}. We say that M_{ij} is a *self-enforcing Nash coalition* in state $\{i,j\}$ if any subcoalition of informed traders, taking the capital supplied by the complement of the subcoalition as given, chooses to abide by the per capita shares assigned by the whole coalition. If this is true for all possible subcoalitions, then the coalition M_{ij} is not subject to collapse

since there is no incentive for any member or group of members to deviate from the proposed M_{ij}.[2] We will refer to this coalition as the "Insider Coalition."

Market clearing will require that the price, say p^*_{ij}, equate the demand for consumption goods with the supply of consumption goods in state $\{i,j\}$:

$$Nw_i p^*_{ij} + M_{ij} p^*_{ij} = S\left(p^*_{ij}\right). \tag{2.8}$$

As before, the supply, $S(p_{ij})$, will be either all the endowments of the late consumers, $N(1 - w_i)e_1$, or some lesser amount if there is storage in equilibrium.

We now define a Nash-type equilibrium in this setting. An *Imperfectly Competitive Rational Expectations Equilibrium* is (a) a price system, $\{p_{ij}\}$, (b) specification of storage strategies for the late consumers, $S\left(p_{ij}\right)$, and (c) a specification of insider coalition strategies, $\{M_{ij}\}$, such that, given $\{p_{ij}\}$, knowledge of the model, and the information set of the informed agents in state $\{i,j\}$, the storage and coalition strategies of the respective agent types are chosen such that (i) their respective *utilities are maximized*, (ii) $\{p_{ij}\}$ *clears the market* in state $\{i,j\}$, and (iii) $\{M_{ij}\}$ is *self-enforcing*.

Let $R^* = q'_h R_h + q'_l R_L$ be the uninformed late consumers' expectation at time 1 of the return on capital when state $\{l,L\}$ actually occurs, where q'_h and q'_l are their posterior probabilities of the states being $w_i = w_h$ and $w_i = w_l$, respectively. The following proposition demonstrates the existence of insider trading by the informed agents.

PROPOSITION 1 (Insider Trading): *Let* $\{\widehat{p}_{ij}\}$ *be the full-information prices for states,* $\{i,j\}$. *If* (i) $e_1(1-w_h)/w_h \leq R^*$ *and* (ii) $\frac{M}{N} \geq \frac{(w_h - w_l)}{(1-w_h)}$, *then there exist Imperfectly Competitive Rational Expectations Equilibrium prices* $\{p^*_{ij}\}$, *where* $p^*_{lH} = \widehat{p}_{lH}$, $p^*_{hL} = \widehat{p}_{hL}$, *and* $p^*_{hH} = p^*_{lL} = \widehat{p}_{hH}$. *That is, these prices are fully revealing in only two of the four states.*

Proof: We will verify that the following specification of prices and strategies constitutes an equilibrium for the assumed parameter values.
State $\{l,H\}$

$$p^*_{lH} = \frac{e_l(1-w_i)}{w_1};\ M_{lH} = 0;\ S\left(p^*_{lH}\right) = N(1-w_l)e_1 \qquad \text{(No Storage)}.$$

2. See Bernheim, Peleg, and Whinston (1987) for the motivation for this definition of a self-enforcing coalition. This equilibrium concept refines the set of possible Nash equilibria of the game between the insiders when they choose the Insider Coalition strategy. For our purposes it focuses attention on equilibria of interest, namely ones in which insider trading occurs.

State $\{h, L\}$

$$p^*_{hL} = R_L;\ M_{hL} = 0;\ S\left(p^*_{hL}\right) < N\left(1 - w_l\right) e_1 \qquad \text{(Some Storage)}.$$

State $\{h, H\}$

$$p^*_{hH} = \frac{e_1(1 - w_h)}{w_h};\ M_{hH} = 0;\ S\left(p^*_{hH}\right) = N\left(1 - w_l\right) e_1 \qquad \text{(No Storage)}.$$

State $\{l, L\}$

$$p^*_{lL} = \frac{e_1(1 - w_h)}{w_h};\ M_{lL} = \frac{N(w_h - w_l)}{(1 - w_h)};\ S\left(p^*_{lL}\right) = N\left(1 - w_h\right) e_1 \quad \text{(No Storage)}.$$

The proposed equilibrium prices in the first three states, $\{l, H\}$, $\{h, L\}$, and $\{h, H\}$, are the full-information prices. In the states $\{l, H\}$ and $\{h, L\}$, prices are fully revealing and are market clearing. It remains, then, to show that the actions of the insider coalition can cause prices to only partially reveal information in the states $\{h, H\}$ and $\{l, L\}$.

In state $\{l, L\}$, the return on the capital goods is low, and informed agents would like to sell their capital goods in exchange for consumption goods at the assumed equilibrium price. They will then store the consumption goods for one period. Since the proportion of the late consumers is low, w_l, the informed coalition can mimic the state $\{h, H\}$ where there are many late consumers and the informed agents don't enter the market.

Thus, if the late consumers supply all their endowment of consumption goods, then market clearing requires

$$N w_l p^*_{iL} + M_{iL} p^*_{iL} = N\, e_1 \left(1 - w_l\right). \tag{2.9}$$

Now, set $p^*_{hH} = p^*_{hH} = e_1 \frac{(1 - w_h)}{w_h}$ and solve for M_{lL} :

$$M_{lL} = \frac{N(w_h - w_l)}{(1 - w_h)}. \tag{2.10}$$

Condition (ii) of the proposition insures that insiders have sufficient capital for (2.10) to hold. By supplying M_{lL} units of the capital good in exchange for the endowment good, the insider coalition can create the false impression that the state is $\{h, H\}$ when, in fact, the state is $\{l, L\}$. However, for this to be successful, two further considerations need to be examined.

First, will late consumers choose to sell their endowment when they see the market clearing price P^*_{lL}? They will if, on average, it is profitable to do so, i.e., when condition (i) of the proposition holds:

$$p^*_{lL} = e_1 \frac{(1 - w_h)}{w_h} \leq R^* = q'_h R_H + q'_l R_L. \tag{2.11}$$

If late consumers form their expectation of the state being $\{l, L\}$ or $\{h, H\}$ in a Bayesian fashion, conditional on the fact that they, themselves, are late consumers, then

$$q_h' = q_h \frac{(1-w_h)}{q_h(1-w_h) + q_l(1-w_l)'}$$

$$q_l' = q_l \frac{(1-w_l)}{q_h(1-w_h) + q_l(1-w_l)\cdot}$$

Condition (2.11) says that, even though late consumers know that the informed coalition will cheat them in state $\{l, L\}$ and that this cannot be detected, still it is optimal to sell all their endowment. It is optimal if q_h' is sufficiently large, so that most often the true (but unobserved) state is $\{h, H\}$.

Second, we must check that M_{lL} is a self-enforcing Nash coalition. If there is a total of M units of capital owned by the informed agents, and they are all in the coalition, then each can exchange M_{lL}/M per unit of the capital for endowment goods. Note that, if any member or group of members independently demands additional endowments, then the market clearing condition (2.9) will not hold at P_{lL}^* and the new price will reveal the collusion. Uninformed agents will infer the truth. If the state $\{l, L\}$ is revealed, late consumers will not be willing to sell their endowments. If there is a deviation from M_{lL}, then the informed agents as a group will not benefit, including the member or group who deviated. Therefore, since any deviation results in a fully revealing price and, hence, no benefits to informed agents, M_{lL} is self-enforcing. Q.E.D.

We can now calculate the expected utility per unit of capital endowment for the informed traders. While M is the total amount of capital endowment of the informed agents, the coalition can only sell M_{lL} units in state $\{l, L\}$. Therefore,

$$\begin{aligned} E(C_2) &= e_2 + \frac{R_H}{2} + \frac{q_h R_L}{2} + \frac{q_l}{2}\left[R_L + w_m(p_{lL}^* - R_L)\right] \\ &= e_2 + \overline{R} + \frac{q_l}{2} w_m (p_{lL}^* - R_L), \end{aligned} \tag{2.12}$$

where $w_m \equiv \frac{M_{lL}}{M} = \frac{N(w_h - w_i)}{M(1-W_h)}$.

Since $R_L < P_{lL}^*$, by assumption (2.5), the expected utility of an informed trader exceeds the full-information expected utility since $w_m > 0$.

Likewise, we can calculate the expected utility of liquidity traders. It is straightforward to show that

$$E[C_1 + C_2] = e_1 + \overline{R} - \frac{q_1}{2}\frac{(w_h - w_l)}{(1-w_h)}\left[p_{lL}^* - R_L\right]. \tag{2.13}$$

Note that this utility is less than that of the full-information case. We now turn to investigating whether the liquidity traders can prevent being victimized by the informed traders.

2.4. PRIVATE LIQUIDITY CREATION

In the previous section, liquidity traders were not allowed to contract. The result was the existence of insider trading that increased the welfare of informed traders at the expense of the liquidity traders. We now allow the liquidity traders to respond by contracting. We show that allowing liquidity traders to contract can prevent insider trading by breaking the informed agents' coalition; i.e., the insider trading equilibrium analyzed in the previous section will no longer exist. Next, we show that an alternative equilibrium characterized by bank intermediation can exist. Finally, we show that the allocation achieved with the bank can be replicated at the firm level with corporations issuing riskless debt.

2.4.1. Bank Intermediation and Liquidity Creation

Suppose at date $t = 0$ the following contract is offered to agents. An intermediary will be set up which pools agents' capital and issues securities to them. Let $A = N_I + M_I$ be the total endowment of the capital good contributed by members of this intermediary as of date $t = 0$, where $N_I = N - N_S$ and $M_I = M - M_S$. The subscript I refers to the capital of agents joining the intermediary, and S refers to the capital of agents continuing to invest in the stock market. The total return of the intermediary's assets at date $t = 2$ is AR_i, $i = H, L$. Ownership of two types of claims on this capital is offered to agents: debt claims and equity claims. Let D and E (whose sum equals A) be the total amount of capital contributed by agents who own debt and equity claims, respectively.

The contract also imposes a debt-to-equity ratio ceiling such that the total payment promised to debt claim, DR_D, must be less than or equal to AR_L; i.e., debt claims are required to be riskless:

$$DR_D \leq AR_L = (D + E) R_L. \tag{2.14}$$

Therefore,

$$\frac{D}{(D+E)} \leq \frac{R_L}{R_D} \text{ or } E \geq \frac{D(R_D - R_L)}{R_L}. \tag{2.15}$$

We would like to consider whether offering agents this intermediary contract would affect the Imperfectly Competitive Rational Expectations Equilibrium analyzed in the previous section. Before stating a series of propositions related to this issue, we make an additional assumption that will simplify the proof of the first of these propositions. We assume that conditional on being a late consumer, the probabilities of the state being $w_i = w_h$ or $w_i = w_l$ are equally likely. If late consumers form expectations in a Bayesian manner, this implies

$$q_h (1 - w_h) = q_l (1 - w_l). \tag{2.16}$$

Now suppose that liquidity traders are allowed to offer the intermediary contract to all agents as a possible trading mechanism. It is clear that, for R_D sufficiently high, liquidity traders are better off holding bank debt. The question is whether the informed agents can be induced to defect from the Insider Coalition to become the bank's equity holders. If this occurs, the intermediary contract will be feasible and the equilibrium of the previous section will not exist.

PROPOSITION 2 (Nonexistence of Stock Market Insider Equilibrium): *Consider a small number of liquidity traders, say N_I (close to zero), choosing to form a bank. Then, if the ratio of informed to uninformed agents' capital, $\frac{M}{N}$, is sufficiently large, there exists a rate of return on intermediary debt, R_D, such that (i) debt is riskless, (ii) liquidity traders prefer to invest their capital in the debt of the intermediary rather than the stock market, and (iii) individual informed agents prefer to invest their capital in the equity of the intermediary rather than the stock market insider coalition.*

Proof: See the Appendix.

Proposition 2 provides a condition under which individual liquidity traders and informed agents have an incentive to form an intermediary at time 0 rather than invest in the stock market. The higher the ratio of total capital of the informed agents to that of the liquidity traders, the smaller will be the expected profits of the informed agents in the Insider Coalition. Therefore, the higher this capital ratio, the smaller is the required rate of return on bank equity necessary to induce an individual insider to join the bank and defect from the Insider Coalition. Consequently, if the required return on bank equity is not too large, the rate of return on bank debt will be large enough to attract an individual liquidity trader away from the stock market as well.

The next proposition states that an equilibrium can exist where *all* liquidity traders choose to purchase the riskless debt of an intermediary and informed agents derive no advantage from operating an Insider Coalition in the stock market. The proof of this proposition assumes the following condition, which includes condition (2.5) assumed previously:

$$R_L < \frac{e_1(1-w_h)}{w_h} < \overline{R} < \frac{e_1(1-w_l)}{w_l} < R_H. \tag{2.17}$$

PROPOSITION 3 (Existence of an Intermediary Equilibrium): *If $\frac{M}{N}$ is sufficiently large, then there exists an equilibrium where* (i) *all liquidity traders purchase riskless debt of the intermediary and* (ii) *informed agents will choose to contribute equity capital.*

Proof: See the Appendix.

The intuition behind this result is that, if informed agents' capital is sufficiently large relative to that of the liquidity traders, it is feasible for a bank to

issue sufficient riskless debt that can be used by all liquidity traders for transactions.[3] Implicitly, the existence of this bank contract allows the informed agents to be identified so that trade with them can be avoided. All liquidity traders who are early consumers will trade bank debt for endowment at date $t = 1$. Late consumers considering selling their endowment at date $t = 1$ will never choose to purchase stock market capital because they know that only informed agents will be supplying stock market capital for endowment, and then only when the return on capital is low. Thus, the stock market becomes an Akerloff (1970) "Lemons" market, and late consumers will choose to trade only with early consumers selling intermediary debt. In this sense, liquidity traders are able to "protect" themselves from possible disadvantageous trades with the better informed agents.

In summary, we have shown that conditions exist where liquidity traders are better off holding intermediary debt which is made riskless because some informed investors will voluntarily contribute equity capital for the intermediary. Under these conditions, with $N_I = N$ and $N_s = 0$, the advantage that the Insider Coalition derives from superior information is completely eliminated. With no one to trade with at date $t = 1$ except other informed agents, informed agents' expected rate of return on stock is simply $\overline{R}$. With sufficient defections from the Insider Coalition, the competitive expected rate of return on intermediary equity will also approach $\overline{R}$, resulting in a deposit rate, R_D, with a limiting value equal to $\overline{R}$. Hence, the private intermediary contract can result in an allocation which gives all agents an expected utility arbitrarily close to the full-information case.

2.4.2. Corporate Debt and Liquidity Creation

So far we have implicitly assumed that "firms" do not issue debt. That is, when we considered the stock market equilibrium in Section 2.3, we imagined individuals exchanging their capital with firms who issued them equity shares. In this section we briefly consider what happens if the firms are willing to buy capital at $t = 0$ in exchange for either debt or equity. So now there exists a market for corporate debt, such as commercial paper.

Suppose a firm offers to pay R_D per dollar of debt and issues an amount of riskless debt such that $DR_D = AR_L$, where $A = D + E$ is the firm's total assets. Then it is immediately apparent that the firm can offer the same riskless debt as the bank intermediary we described previously. All of the above arguments about

3. In addition, as is shown in the Appendix, the greater R_L is, the higher is the feasible leverage of the intermediary, i.e., the smaller is the proportion of informed agents needed to join the intermediary to make its debt riskless. The greater the leverage, the less R_D needs to be lowered in order to raise the expected rate of return on the intermediary's equity in order to attract informed agents.

the bank now apply to the firm. Agents need not directly hold the claims of firms, but mutual funds could arise to specialize in holding either debt or equity claims. In particular, funds similar to money market mutual funds could purchase the high-grade debt (e.g., commercial paper) of firms. As before, the equilibrium would be for all liquidity traders to buy claims on the debt fund and all informed traders to buy claims on the firm's equity. We comment further on this in our concluding remarks.

2.5. DEPOSIT INSURANCE AND A GOVERNMENT DEBT MARKET

A deposit insurance system for banks can also satisfy the liquidity traders' desire for a safe asset for trading. In this section we show how deposit insurance can replicate the allocation of the previous section when intermediary debt is risky. In addition, we show that development of a government debt market is similar to deposit insurance, as it involves government creation of a risk-free asset. In a like manner, a government debt market can replicate the riskless corporate debt contract when riskless corporate debt is in insufficient supply. The government can succeed where private contracting fails due to its ability to enforce lump sum taxation. It is the revenue from this taxation that accounts for the government's ability to create riskless securities.

As Merton (1977) has observed, "the traditional advantages to depositors of using a bank rather than making direct market purchases of fixed-income securities . . . economies of scale, smaller transactions costs, liquidity, and convenience . . . are only important advantages if deposits can be treated as riskless." Presumably, if deposits were not riskless, then small agents would face information and surveillance costs necessary to evaluate the current risk of bank liabilities. Without this information, other informed agents might then take advantage of them. Consequently, less informed agents would benefit if there were deposit insurance. Indeed, a stated goal of government deposit insurance is to protect the small investor.

Suppose that deposits are risky, i.e., $DR_D > AR_L$. This would be the case if, for example, the capital endowments of the informed agents were too small to provide enough riskless debt or if $R_L = 0$. In other words, if the low return rate state of the world is realized, then deposits will incur a capital loss. The insurance system works as follows. If R_L is realized, so that the bank would not be able to meet its promised payments at time $t = 2$, then the government is assumed to tax all late consuming agents in proportion to their endowment in order to raise enough revenue to pay off the bank debt at par.[4] The government will also charge

4. The government is assumed to observe the bank failure at date $t = 2$.

an insurance premium at time $t = 2$ that the bank pays if it does not fail, i.e., when R_H is realized, which is allocated to all late consuming agents.

Let T be the tax revenue collected when the bank fails. In order to avoid a capital loss on deposits if R_L is realized, the amount of insurance needed is $T = DR_D - AR_L$. Each agent consuming at date $t = 2$ pays a share of T. At $t = 2$ there are informed agents who were endowed with M units of capital and $N(1 - w_i)$, $w_i = w_l$ or w_h, late consuming liquidity traders, each having been endowed with one unit of capital. This insurance arrangement will only be feasible if, regardless of the proportion of early consumers, the remaining agents can afford to pay the tax. Thus, feasibility requires

$$T \big/ [M + N(1 - w_i)] < e_2, \quad i = l, h, \tag{2.18a}$$

$$T \big/ [M + N(1 - w_i)] < R_D + e_1 \frac{R_D}{P_{Di}}, \quad i = l, h. \tag{2.18b}$$

Informed agents have, at least, Me_2, their second-period endowment.[5] Thus, the tax per unit capital cannot exceed the e_2 endowment. This is requirement (2.18a) above. Similarly, (2.18b) requires that the late consuming liquidity traders, who have assets of $R_D + e_1 \frac{R_D}{p_{Di}}$, be able to afford the tax. (The values of p_{Di} are given by (2A.18) in the Appendix.)

If the bank does not fail, then the bank pays an insurance premium of ϕ to the rest of the economy, which consists of all informed agents and depositors. The expected return to the bank equity holders in the presence of deposit insurance is

$$E[R_E]E = (1/2)[R_H(D + E) - (R_D + \phi)D] + (1/2) \cdot 0. \tag{2.19}$$

It is straightforward to solve for a fair insurance premium. Since bank failure and bank solvency are equally likely, i.e., R_L and R_H each occur with probability one half, a fair insurance premium equates the amount paid as a premium in the high state with the amount of insurance coverage in the low state:

$$\phi D = T = DR_D - (D + E)R_L, \tag{2.20}$$

which implies that

$$\phi = R_D - \frac{(D + E)}{D} R_L. \tag{2.21}$$

Substituting (2.21), the expression for the fair deposit insurance premium, into (2.19) yields

$$E[R_E]E = R(D + E) - R_D D. \tag{2.22}$$

5. Informed agents holding bank equity have only e_2 per unit of initial endowed capital since their bank equity is worthless, while informed agents in the stock market have $e_2 + R_L$.

As in the previous section, consider a competitive equilibrium where the expected rate of return on equity approaches $\overline{R}$. In this case, equation (2.22) shows that R_D also approaches $\overline{R}$. Therefore, the allocation under the deposit insurance scheme gives agents the same expected utility as in the case of the private uninsured intermediary considered in the previous section. In summary, we have shown the following.

PROPOSITION 4 (Deposit Insurance): *When bank debt is risky, the tax-subsidy scheme* $\{T, \phi\}$*, defined above, can implement an allocation which gives all agents the same expected utility as in the riskless private bank deposit allocation.*

Similar to government intervention as a deposit insurer, we can consider whether government intervention can benefit uninformed agents when firms issue corporate debt, as was described previously. Let us start from the assumption that each firm issues riskless debt such that

$$A_i R_L \geq D_i R_D, \tag{2.23}$$

where A_i and D_i are the assets and debt of firm i, respectively. However, suppose that the assets of firms are of sufficient risk to preclude uninformed agents from placing their entire wealth in risk-free corporate debt. In this case, government intervention in the form of a government debt market can allow uninformed agents to replicate the allocation of the previous Section 2.4.2, where riskless corporate debt was in sufficient supply.

As with the deposit insurance scheme, the government can create additional two-period risk-free securities backed by lump sum taxation of agents' endowment in period 2. The government simply issues claims on second-period endowment equal to the difference between uniformed agents' time 0 endowment and the supply of risk-free corporate debt, so that the government sells bonds for capital equal to $N - D$ at time 0. Since government and firm debt are perfect substitutes, they each pay a two-period return of R_D, implying that the time $t = 2$ maturity value of government bonds B equals

$$B = (N - D)R_D. \tag{2.24}$$

The government is assumed to invest the capital it acquires at time 0, either directly investing it itself or giving it to firms which issue it equity shares. At time $t = 2$, this investment is worth $(N - D)R_i$, $i = H, L$. The short fall (excess) between this investment return and the promised payments on bonds, B, is made up by lump sum taxation (subsidization) of late consumers, subject to feasibility requirements similar to (2.18a) and (2.18b). Competitive equilibrium implies that the expected return on equity as well as the return on riskless debt will equal $\overline{R}$.

Thus, the additional debt supplied by the government can allow uninformed agents to purchase sufficient risk-free securities to meet their demands for liquidity. Hence, this intervention can also restore for the uninformed agents an allocation which gives them the same expected utility as in the full-information case.

2.6. CONCLUSION

The historically popular notion that informed agents can benefit at the expense of uninformed agents is true in the setting which we have analyzed. Informed agents can form an insider coalition which is self-enforcing and can benefit at the expense of the lesser informed agents. When this condition exists, a demand for liquid securities by uninformed agents will result. By splitting risky cash flows, liquid securities are created which have the effect of eliminating the potential advantage possessed by better informed agents.

Liquidity can be created through the formation of banks. We have formalized a traditional rationale for the existence of banks and deposit insurance, namely that they provide a riskless transactions medium that eliminates the need of uninformed agents to trade in assets whose returns are known by better informed agents. By issuing deposits, banks create "riskless" securities for trading purposes. In instances where bank asset risk is such that uninsured deposits cannot be made riskless, we have shown that deposit insurance can replicate the allocation achieved with riskless private bank deposits.[6] In addition, liquid securities can also be created through the formation of corporate debt or government securities markets. As an alternative to bank intermediation, firms can split risky cash flows, thereby creating a safer security (debt).

An empirical implication of our model is that transactions securities should be the most actively traded assets. This is consistent with the relatively high turnover in ownership of insured bank liabilities and Treasury securities. Corporate debt, on the other hand, is much less actively traded, suggesting that our assumption that firms can create riskless securities simply by splitting the cash flows of their underlying assets is not completely accurate.

For tractability, we studied a model with a single source of asset risk. Clearly, with multiple sources of asset risk, diversification would provide another, perhaps complementary, channel for the reduction of risk. This channel implies

6. An issue which we have not considered concerns possible equilibria where banks exist but their uninsured bank deposits are risky. In this situation we conjecture that the liquidity traders would be better off than without the bank but clearly would not be as well off as the case of riskless bank debt. The value of risky bank debt would depend on the state of nature, but to a lesser extent than would stock. Informed traders might still use their information advantage.

combining imperfectly correlated assets to reduce risk, rather than splitting cash flows. These issues are investigated in Gorton and Pennacchi (1989). The creation of mutual funds holding a diversified portfolio of corporate debt can alleviate the inability of individual firms to create riskless debt. For example, money market mutual funds are large holders of commercial paper, and the shares of these funds provide a potentially important transactions medium.[7]

Due to the recent growth of the market for short-term corporate debt, the possibility of substituting money market mutual fund shares for bank debt is intriguing.[8] A public policy debate has smoldered around whether such alternative instruments should be encouraged or restricted as transactions media. In our analysis there is not reason to prefer bank debt over money market mutual funds. However, extending our analysis to consider the regulatory distortions and monitoring costs associated with bank deposit insurance might lead to a preference for a money market mutual fund-based transactions system.

APPENDIX

Proof of Proposition 2: Step 1 of the proof is to consider the situation of the liquidity traders. Given the feasibility of the intermediary, we derive the conditions under which they are better off purchasing the intermediary's debt rather than investing their capital in the stock market. Step 2 considers the informed agents and shows that, under the conditions derived in step 1, they may be individually better off by becoming equity holders in the intermediary rather than being members of the Insider Coalition that operates in the stock market. Thus, if informed agents are willing to contribute equity capital, the intermediary contract is feasible.

Step 1: Let p_{Dij} be the number of endowment units received in exchange for one unit of the debt claim at date $t = 1$ when the state is $\{i, j\}$, where $i = l, h$, and $j = L, H$. Because of the risk neutrality of uninformed agents, at time $t = 1$ it must be the case that

$$\frac{R_D}{p_{Dij}} = \frac{R^e}{p_{ij}} \equiv r_{ij}, \tag{2A.1}$$

7. Currently, the transactions services provided by money market mutual fund shares may by inhibited by regulation which denies these mutual funds independent access to the payments system. Money market mutual fund check and wire transfers must be carried out through commercial banks.

8. Perhaps an unplanned benefit of large government budget deficits has been an increased supply of riskless debt, further adding to the feasibility of a transactions system backed by money market instruments.

where R^e is the uninformed late consumers' expectation at time $t = 1$ of the return on the capital good at time $t = 2$ and r_{ij} is defined to be this common expected reinvestment rate when state $\{i,j\}$ occurs.

We can now calculate the time $t = 0$ expected utility of an uninformed agent who invests capital in the stock market, $E_s[C_1 + C_2]$, and the utility of an uninformed agent who invests capital in the debt of the intermediary, $E_I[C + C_2]$

$$E_s[C_1 + C_2] = \sum_{\{i,j\}} \frac{q_i}{2}\left(w_i\left(e_1 + p_{ij}\right) + (1 - w_i)\, r_{ij}\left(e_1 + p_{ij}\right)\right), \qquad \textbf{(2A.2)}$$

$$E_I[C_1 + C_2] = \sum_{\{i,j\}} \frac{q_i}{2}\left(w_i\left(e_1 + p_{Dij}\right) + (1 - w_i)\, r_{ij}\left(e_1 + p_{Dij}\right)\right). \qquad \textbf{(2A.3)}$$

The difference between (2A.3) and (2A.2) will determine whether uninformed agents have an incentive to invest in the intermediary.

$$E_I[C_1 + C_2] - E_S[C_1 + C_2] = \sum_{\{i,j\}} \frac{q_i}{2}\left(p_{Dij} - p_{ij}\right)\left(w_i + (1 - w_i)\, r_{ij}\right). \quad \textbf{(2A.4)}$$

To determine the sign of (2A.4), we need to compute the prices p_{Dij} and p_{ij}. As in Section 2.3 of the text, these prices will, in general, depend on the parameters of the model as well as the actions of the informed agents. Analogous to condition (2.5) in the text, we state the following conditions:

$$R_H > \frac{e_1(1 - w_l)}{w_l} + \frac{N_I}{N_S}\left(\frac{e_1(1 - w_l)}{w_l} - R_D\right), \qquad \textbf{(2A.5)}$$

$$R_L < \frac{e_1(1 - w_h)}{w_h} + \frac{N_I}{N_S}\left(\frac{e_1(1 - w_h)}{w_h} - R_D\right). \qquad \textbf{(2A.6)}$$

Note that, for N_I sufficiently small relative to Ns, conditions (2A.5) and (2A.6) will hold if condition (2.5) holds. Thus, we wish to examine the incentives for a small group of uninformed agents to join an intermediary, given that there currently exists a large number in the stock market.

Analogous to the results of Section 2.3, if conditions (2A.5) and (2A.6) hold, then states $\{l,H\}$ and $\{h,L\}$ are fully revealing, while an Insider Coalition can form to purchase endowment in state $\{l,L\}$ to mimic the prices of all securities in state $\{h,H\}$. Using (A1) and equating demands and supply of the endowment good lead to the following set of state-contingent prices and time $t = 1$ reinvestment rates:

$$\text{(No Storage)} \qquad p_{Dl|H} = \frac{e_1(1 - w_l)NR_D}{w_l(N_I R_D + N_s R_H)}, P_{l|H} = P_{Dl|H}\frac{R_H}{R_D}, r_{l|H} = \frac{R_D}{P_{Dl|H}},$$

(Storage) $\qquad P_{DhL} = R_D, P_{hL} = R_L, r_{hL} = 1,$

(No Storage)

$$P_{DlL} = P_{DhH} = \frac{e_1(1-w_h)NR_D}{w_h(N_lR_D + N_SR^*)}, P_{lL} = P_{hH} = P_{DlL} = \frac{R^*}{R_D},$$

$$r_{lL} = r_{hH} = \frac{R_D}{P_{DlL}}, \tag{2A.7}$$

where R^* is the late consumers' expectation at time 1 of the return on capital at date 2, when the state is only partially revealed to be either $\{l, L\}$ or $\{h, H\}$. The formula for R^* is given in equation (2.11) of the text. Substituting these prices and reinvestment rates into (2A.4) and simplifying, one obtains

$$\begin{aligned} &E_I[C_1 + C_2] - E_s[C_1 + C_2] = \\ &{}^1\!/_2\,(R_D - R^*)\left[\frac{(1-w_h)\,N\,e_1\left(q_hw_h + q_lw_l\right)}{w_h\,(N_IR_D + N_SR^*)} + q_h\,(1-w_h) + q_l\,(1-w_l)\right] \\ &\quad + {}^1\!/_2(R_D - R_H)q_l\,(1-w_l)\left[\frac{N\,e_1}{N_IR_D + N_SR_H} + 1\right] + {}^1\!/_2\,(R_D - R_L)q_h. \end{aligned} \tag{2A.8}$$

It is straightforward to verify that (2A.8) is a strictly increasing function of R_D and, for R_D sufficiently large, uninformed agents will prefer joining the intermediary. Furthermore, we can also show that there exists a value of $R_D < \overline{R}$ for which (2A.8) will be positive when all uninformed investors initially invest in the stock market, i.e., when N_l is small. Taking the limit as $N_I \to 0$ $(or\ N_S \to N)$,

$$\begin{aligned} &\lim_{N_I \to 0} E_I[C_1 + C_2] - E_S[C_1 + C_2] = \\ &\quad {}^1\!/_2\,(R_D - R^*)\left[\frac{(1-w_h)}{R^*w_h}e_1\left(q_hw_h + q_lw_l\right) + q_h\,(1-w_h) + q_l\,(1-w_l)\right] \\ &\quad + {}^1\!/_2\,(R_D - R_H)\,q_l\,(1-w_l)\left(\frac{e_1}{R_H} + 1\right) + {}^1\!/_2\,(R_D - R_L)\,q_h. \end{aligned} \tag{2A.9}$$

Setting the right-hand side of equation (2A.9) to zero, we can solve for the minimum return on intermediary debt, R_D^m, for which uninformed agents are as well off joining the intermediary as they are staying in the stock market. For the simplifying case of condition (2.16), that, conditional on being a late consumer, the probability of the state being h or l is equally likely $(R^* = \overline{R})$, we have

$$R_D^m = \overline{R} - q_l(1-w_l)\left(\frac{R_H - R_L}{2}\right)\left[\frac{w_h}{1-w_h} - \frac{e_1}{R_H}\right]/\theta, \tag{2A.10}$$

where

$$\theta \equiv \left[\frac{(1-w_h)}{w_h}\,\frac{e_1}{\overline{R}}(q_h w_h + q_l w_l) + q_l(1-w_l)\left(3+\frac{e_1}{R_H}\right) + q_h\right] > 0.$$

The term in brackets on the right-hand side of (2A.10) is strictly positive because of condition (2.5). Since (2A.9) is continuous and strictly increasing in R_D, it must also be strictly positive for some value of R_D less than $\overline{R}$.

Step 2: Given that liquidity traders have an incentive to leave the stock market and join the intermediary for $R_D > R_D^m$, we now show that the intermediary contract will be feasible if informed agents can be induced to provide equity financing rather than invest their capital with the stock market Insider Coalition.

The informed agents who are members of the stock market Insider Coalition will sell their capital to mimic the state $\{h, H\}$ when the state is actually $\{l, L\}$. They purchase endowment in the amount:

$$M_{lL} = \frac{(w_h - w_l)}{(1-w_h)}\left(N_s + N_I R_D/R^*\right), \qquad \textbf{(2A.11)}$$

which results in their time 0 expected utility per unit capital being

$$E[C_2] = e_2 + \overline{R} + \frac{q_l}{2}\frac{M_{lL}}{M_S}\left(p_{lL} - R_L\right), \qquad \textbf{(2A.12)}$$

where p_{lL} is given by (2A.7).

Note that, for $R_D < R^*$, $M_{il}$is less than its value for the case where $N_I = 0$, which was analyzed in Section 2.3, while $p^*{}_{lL}$ is less than p^*_{lL} given in Section 2.3. Thus, the expected utility of the informed agents falls in this case if M_S stays the same. Now if some informed agents defect from the Insider Coalition and invest their capital, equal to M_I, in the equity of the intermediary, their expected return will be

$$E[M_I R_E] = \overline{R}(N_I + M_I) - R_D N_I. \qquad \textbf{(2A.13)}$$

If the intermediary's capital constraint is binding so that N_I and M_I follow the debt and equity proportions given in equation (2.15), then the expected return on intermediary equity equals

$$E[R_E] = \overline{R} + \left(\overline{R} - R_D\right)\tfrac{R_L}{(R_D - R_L)}. \qquad \textbf{(2A.14)}$$

Thus, comparing (2A.14) with (2A.12), we see that an informed agent who invests in the equity of the intermediary will have a higher expected return than an informed agent in the Insider Coalition if

$$\left(\frac{\overline{R} - R_D}{R_D - R_L}\right) R_L > \frac{q_l}{2}\frac{(N_S + N_I R_D/R^*)}{M_S}\frac{\left(w_h - w_l\right)}{\left(1-w_h\right)}\left(p_{lL} - R_L\right). \qquad \textbf{(2A.15)}$$

Consider the incentive for informed investors to defect from the stock market coalition when initially N_S is close to N. Taking the limit of (2A.15) as N_I goes to zero and rearranging terms result in

$$\left(\overline{R} - R_D\right) > \frac{q_l}{2}\frac{N}{M}\frac{\left(w_h - w_l\right)}{\left(1 - w_h\right)}\left[\frac{e_1\left(1 - w_h\right)}{w_h R_L} - 1\right]\left(R_D - R_L\right). \qquad \textbf{(2A.16)}$$

Now suppose R_D is set such that $\overline{R} > R_D \geq R_D^m$, where R_D^m is given by (2A.10). Then both sides of condition (2A.16) are strictly positive, but the right-hand side of (2A.16) can be made sufficiently small for M sufficiently large. (Note that R_D^m is independent of M.) Thus, for M/N sufficiently large, a return on intermediary debt can be offered which gives both uninformed and informed agents the incentive to start an intermediary.

Proof of Proposition 3: We first take the feasibility of the intermediary for $N_I = N$ as given and later show that this holds for M/N sufficiently large. If all liquidity traders initially invest in the riskless debt of the intermediary, consider the possibility of the informed traders being able to strategically purchase the endowment of the late consumers when the return on stock market capital is low.

Given condition (2.17), consider a return on intermediary debt, R_D, such that

$$e_1\frac{(1 - w_h)}{e_1 w_h} < R_D \leq \overline{R}. \qquad \textbf{(2A.17)}$$

Similar to the analysis of Section 2.2 in the text, it is straightforward to show that a full-information equilibrium would result in the time $t = 1$ prices of intermediary debt equal to

$$\text{(Some Storage)} \qquad p_{Dlj} = R_D, \quad j = L, H,$$

$$\text{(No Storage)} \qquad p_{Dhj} = e_1\left(1 - w_h\right)/w_h, \quad j = L, H. \qquad \textbf{(2A.18)}$$

In other words, some storage occurs whenever there is a low proportion, w_l, of early consumers, and no storage occurs whenever there is a high proportion, w_h, of early consumers. In equilibrium, the price of stock market capital will satisfy

$$p_{ij} = p_{Dij}E\left[R_j\right]/R_D = p_{Dij}\frac{R_j}{R_D}. \qquad \textbf{(2A.19)}$$

Now consider the case of asymmetric information. Stock market insiders would like to be able to purchase endowment and sell stock market capital at time 1 when the return on capital is low, R_L. Potentially, they could do this, as before, when state $\{l, L\}$ occurs, by purchasing endowment from late consumers. However, rational late consumers would never choose to sell their endowment for stock market capital because the only sellers of stock market capital are

informed agents, who the late consumers know would only choose to sell capital when the return is R_L. Unlike the situation considered in Section 2.3, where liquidity traders invested in the stock market at time 0, late consumers will now realize that they will only be trading capital with informed agents, and then only when the return on capital is R_L. Hence, late consumers will only offer a price for stock market capital of

$$p_{ij} = p_{Dij}\frac{R_L}{R_D}. \tag{2A.20}$$

At this price, there would be no incentive for informed agents to purchase endowment. Since late consumers would only sell endowment for the riskless debt of early consumers, p_{Dij} would always be equal to its full-information price given in (2A.18). This results in the expected utility of uninformed agents being equal to

$$E[C_1 + C_2] = e_1 + R_D \tag{2A.21}$$

and the stock market Insider Coalition being devoid of power, their return on capital simply being equal to $\overline{R}$. Hence, in order to attract informed agents to contribute to the intermediary, R_D need only be an arbitrarily small amount less than $\overline{R}$, and uninformed agents' utility would approach their full-information level. In addition, it is straightforward to show that individual liquidity traders would never choose to invest their capital in the stock market rather than the intermediary since, if they turn out to be an early consumer, they can only sell their capital to late consumers at a price which always reflects the return on capital being R_L given by (2A.20).

Finally, to show that this equilibrium is feasible, informed agents must have sufficient capital in order to purchase the minimum amount of intermediary equity required to make the intermediary's debt riskless. Using condition (2.15), with $D = N$ we have

$$\frac{M}{N} > \frac{(R_D - R_L)}{R_L}. \tag{2A.22}$$

Note that the larger R_L is, the smaller is the amount of equity capital needed to enable the intermediary's debt to be riskless.

REFERENCES

Akerloff, G., 1970, The market for lemons: Qualitative uncertainty and the market mechanism, *Quarterly Journal of Economics* 84, 488–500.

Bernheim, B. Douglas, Bezalel Peleg, and Michael Whinston, 1987, Coalition-proof Nash equilibria, *Journal of Economic Theory* 42, 1–12.

Campbell, T. and W. Kracaw, 1980, Information production, market signalling and the theory of financial intermediation, *Journal of Finance* 35, 863–81.

Cone, Kenneth, 1983, *Regulation of Depository Institutions*, Ph.D. Thesis, Stanford University.

Diamond, D., 1984, Financial intermediation and delegated monitoring, *Review of Economic Studies* 51.

Diamond, D., and P. Dybvig, 1983, Bank runs, liquidity and deposit-insurance, *Journal of Political Economy* 91, 401–19.

Gorton, Gary B. and George Pennacchi, 1989, Security baskets and index-linked securities, Working paper, The Wharton School, University of Pennsylvania.

Grinblatt, Mark S. and Stephen A. Ross, 1985, Market power in a securities market with endogenous information, *Quarterly Journal of Economics*, 1143–167.

Jacklin, Charles, 1987, Demand deposits, trading restrictions, and risk sharing, in Edward D. Prescott and Neil Wallace, eds.: *Contractual Arrangements for Intertemporal Trade* (University of Minnesota Press, Minneapolis).

———, 1988, Demand equity and deposit insurance, Mimeo, Stanford University.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–36.

Merton, Robert C., 1977, An analytic derivation of the cost of deposit insurance and loan guarantees, *Journal of Banking and Finance* 1, 3–11.

Myers, Stewart and Nicholas Majluf, 1984, Corporate financing and investment decisions when firms have information that investors do not have, *Journal of Financial Economics* 13, 187–222.

Rock, Kevin, 1986, Why new issues are underpriced, *Journal of Financial Economics* 15, 187–212.

3

Reputation Formation in Early Bank Note Markets

GARY B. GORTON*

3.1. INTRODUCTION

This essay studies the formation of reputations in debt markets. It focuses particularly on the market for bank notes during the American Free Banking Era (1838–60), during which time large numbers of firms entered banking and issued debt in the form of perpetual, non-interest-bearing, risky debt claims, offering the right of redemption on demand at par in specie. The consensus of recent research holds that wildcat banking was not a pervasive problem during this period (see Rockoff 1971, 1974, 1975, 1985, 1989; Rolnick and Weber 1982, 1983, 1984, 1988), but there is no explanation of the mechanism that prevented wildcat banking.[1] The main question addressed in this paper is whether

* Thanks to Charles Calomiris, Douglas Diamond, Eugene Fama, David Galenson, Bruce Grundy. Sam Orez, Robert Vishny, two anonymous referees, and participants in seminars at Chicago, Illinois, Northwestern, and the Penn Macro Lunch Group for discussion and suggestions. The research assistance of Sung-ho Ahn, Chip Bayers, Eileen Brenan, Lalit Das, Molly Dooher, Lori Gorton, Henry Kahwaty, Arvind Krishnamurthy, Charles Chao Lim, Robin Pal, Gary Stein, Kok-Hom Teo, and Peter Winkelman was greatly appreciated. This research was supported by National Science Foundation grant SES86-18130 and a University of Pennsylvania Research Fund grant, for which 1 am very grateful.

1. In general, a "wildcat" bank refers to a bank that inflated its currency to the point at which it could not be continuously redeemed. A number of more precise definitions of wildcat banking have been proposed in the literature. Rockoff (1974, 1975) provided the definition that seems to have become standard. According to Rockoff, a necessary condition for wildcat banking was the possibility that free banks could value the bonds backing their note issuance at par when, in fact, the market value was much lower than par. Then a wildcat bank, according to Rockoff (1975), was a bank that deposited backing securities, which were valued at par by the state banking authorities

market mechanisms, monitoring via note redemption and reputation formation, existed that provided incentives for banks not to engage in wildcat banking.

The theory of reputation formation in debt markets that is tested here is due to Diamond (1989). He considers an observationally equivalent cohort of firms issuing debt for the first time. Some of the firms have high-risk, negative net present value, projects; some have low-risk, positive net present value, projects; and some may choose between the high- and low-risk projects. When these firms issue debt for the first time, there is a lemons problem causing lenders to charge a premium to the new firms above the interest rate charged to firms that have credit histories but are otherwise identical (hereafter called "seasoned" firms). Diamond's main result concerns the dynamic behavior of this lemons premium. Over time, lenders observe defaults and, as a consequence, reduce the premium required on the remaining new firms' debt since, on average, firms with high-risk projects will have defaulted. Since, for a given project, the lower interest rate increases the present value of the borrower's rents, the credit history of being a surviving firm is a valuable asset and corresponds to a "reputation." But the lower interest rate has an additional effect as well since the firms that can choose between projects may find the safer project more attractive. The importance of reputation in Diamond's model is that it affects the actions of some borrowers since it is increasingly costly to default as time goes by. Insofar as some borrowers default over time, this incentive becomes stronger.

The theory predicts that (1) firms issuing debt for the first time should pay higher interest rates than otherwise identical firms and (2) over time lenders will lower the premium, conditional on having observed defaults, until, eventually, the premium disappears. This study is concerned with testing these predictions. The predictions of the model are tested in an environment in which the issues arise in a very clear way and that has the advantage of relative ease in testing the theory.

During the American Free Banking Era, many states passed free banking laws that eased the restrictions on entry into banking (see Rockoff 1975; Rolnick and Weber 1983). Banks during this period issued debt primarily in the form of bank notes, which were used as media of exchange. These notes circulated at discounts from face value at locations some distance from the issuing banks. An important issue concerning the period is whether or not some banks had an incentive to behave as "wildcat" banks, that is, banks that chose to inflate their currencies beyond the point at which they could be continuously redeemed, absconding with the proceeds. More generally, the question concerns how well private money systems can function. In particular, does the notion of reputation provide

but, in fact, were worth less than par. Backing its note issue with overvalued securities then allowed this bank to issue notes that were insufficiently backed. The difference was earned as seigniorage, and the bank was left to fail. See Dillistin (1949) for a discussion of the origin of the term.

an effective mechanism for private money-issuing firms not to behave as wildcat banks? Klein (1974) explicitly argues that competitively supplied private monies can exist because of the ability of issuers to establish reputations. The period is also interesting because of enormous technological change: both the railroad and the telegraph were introduced during this period and rapidly diffused across the country. Part of this study assesses the effects of this technological change on reputation formation.

3.1.1. Reputations and Debt Markets

It is not obvious that debt markets behave in the way Diamond hypothesized. It may be the case that there is enough information available initially to discriminate among different types of firms.[2] A related issue concerns which firms choose to issue debt. In the model of Diamond (1989), firms do not have a choice concerning whether to issue debt publicly. The theory may hold if all firms had to issue debt publicly, but, in reality, it may be that firms issue debt only if their characteristics are sufficiently well known that they do not have to pay a premium (relative to seasoned firms) on their initial debt issuance. Alternative sources of borrowing include privately placed debt, bank debt, and venture capitalists. Diamond (1991) considers the same model as in Diamond (1989), but firms have a choice of financing their projects with bank loans or with publicly issued debt. The main result there is that firms will choose to be monitored by banks until their reputations are established and then issue public debt.

The main problem in empirically testing for the presence of reputation effects is that a counterfactual is posed: whether new firms are charged a premium that declines over time requires knowing what the interest rate would be if the same firm had a reputation.[3] Such a comparison poses the difficult problem of finding

2. In the modern era, corporate debt is typically rated before it is issued. Before firms issue debt publicly for the first time, they have credit histories based on experiences with banks and venture capitalists. Using these histories and other publicly available information, ratings firms and market participants may be able to screen borrowers initially so that there is no initial premium charged on their debt and no subsequent learning. Even the category of "no rating" may be informative. The existence of ratings per se is not evidence against the theory since ratings can be subsequently adjusted on the basis of performance.

3. There is a literature that examines the "seasoning process" for corporate bonds, i.e., the differences in yield to maturity between newly issued bonds and bonds that have been outstanding for some time. The most recent results do not seem to find that new issues have higher yields that persist for any significant period (see Ederington 1974; Lindvall 1977; Weinstein 1978; Sorensen 1982; Fung and Rudd 1986; Wasserfallen and Wydler 1988). None of these studies analyzes price differences between bonds that are the obligations of firms issuing debt for the first time and those of experienced or seasoned firms.

a seasoned cohort of firms with the same asset risk.[4] Also, if Diamond (1991) is correct, then new firms borrow from banks and the interest rates on their loans must be compared to the benchmark cohort (but bank loan interest rate data are generally unavailable).

An additional empirical problem is that in Diamond's theory, lenders learn by observing defaults, but they happen only over relatively long periods of time for most entering cohorts of firms. For example, in a study of junk bonds, Asquith, Mullins, and Wolff (1989) find that default rates are low immediately after issue and rise over time. The length of time required for a significant number of defaults to occur, and hence result in a lower premium for the remaining firms, would seem to make tracing the evolution of the premium especially hard since it is measured relative to the fundamentals of an otherwise identical firm. But the longer the time it takes learning to occur, the more likely the fundamentals of the new firms and those of the benchmark firms will diverge.[5]

3.1.2. The Free Banking Era

In this paper a sample of firms that avoids many of the difficulties noted above is used to test the theory. The sample consists of pre–Civil War banks issuing debt for the first time. The debt consists of bank notes, which were the distinct private monies of banks during this period (1838–63). A bank note was a non-interest-bearing, risky, perpetual debt claim on the bank that could be redeemed at par on demand. This was effectively the only private debt that was publicly issued during this period.[6] Moreover, operating as a bank required issuing bank notes. Thus there can be no selection bias in which firms issued debt. All firms operating as banks issued bank notes. Section 3.2 provides some background on bank notes during the Free Banking Era.

To address the issue of why this period was not characterized by widespread overissuance of private money, I focus on four issues. First, I ask whether Diamond's dynamic lemons premium theory characterizes note issuance during

4. Note that this cannot easily be done on the basis of bond ratings. If the Diamond (1989) theory is correct, then new firms should have lower bond ratings than otherwise identical firms. Over time the ratings of the survivors should converge to the rating of the seasoned firms. But, in that case, the benchmark cohort cannot be formed using bond ratings. Since bond ratings are presumably formed using the available information, it is not clear how the researcher, using the same information, can separate risk due to lack of credit history from risk due to fundamentals.

5. The problem may be compounded by the fact that firms issuing for the first time are usually young, smaller, firms—perhaps riskier. A decline in the interest rate may not reflect learning, but changes in the risk of the firm. Young firms have no natural comparison group.

6. In the latter part of the period, railroads issued bonds.

this period; that is, were new banks charged higher premia (relative to otherwise identical seasoned peers), and did these premia decline over time? During the Free Banking Era, bank notes were not rated, and banks could not have prior histories without having issued bank notes. Nevertheless, there may be prior information that is relevant, perhaps concerning the individuals in charge of the bank or information concerning the bank's capital ratio, ratio of notes to capital, amounts of reserves, and so on. The hypotheses are not mutually exclusive: lenders may be able to discriminate to some extent, but reputations may also be important.

The second issue concerns how note holders monitored banks. I show that the redemption option in the bank note contract provided a mechanism for note holders to monitor banks and that a higher discount (from face value) on the notes of new banks would give them an incentive to monitor. By redeeming the notes of new banks with high discounts and observing whether these banks can honor their obligations, note holders learn whether new banks are riskier than other banks at that location.

The third question concerns the effects of cross-section variation in public and private arrangements concerning banking in the various states. If there is sufficiently widespread adverse selection initially, then according to Diamond, reputation cannot serve to deter firms from choosing excessively risky projects, which, in this case, might include banks that "overissue" monies, so-called wildcat banks. The degree of adverse selection may have varied across states, affecting the extent to which the notes of new banks may have been discounted (relative to seasoned peers). I test for the presence of such factors.

Finally, the ability of market participants to produce and receive information about new banks and their ability to exercise the redemption option by carrying the note back to the issuing bank are influenced by technology. In the 1840s the technology available to transmit information and the transportation technology were primitive. But the technology rapidly improved over the period with the introduction of the telegraph and the diffusion of the railroad. I investigate whether technological change affects reputation formation and monitoring using a measure of technological change constructed from pre–Civil War travelers' guides.

3.1.3. Outline of the Argument and Tests

The basic empirical strategy of the paper is to compare the discounts (from face value) on the bank notes of new banks to the discounts on the notes of existing banks with credit histories (seasoned banks) at the same location. But this is reasonable only if the seasoned banks at the particular location are comparable in every way except that they have credit histories. Section 3.3 addresses this issue;

I argue that the notes of all (solvent) seasoned banks (at a given location) will trade at the same price. It is important to establish a priori that all the seasoned banks at a given location have the same expected risk so that the seasoned note prices can be used as benchmarks against which the prices of new banks' notes can be compared. Appendix A presents a model to make this point formally.

The argument depends on showing that the value of a note declines as it is carried further and further away from the issuing bank. This decline in value is greater if the risk of the bank's portfolio is greater. A consequence is that consumers are not indifferent between the notes of two banks an equal distance away, but with different risks, even if those risks are priced. The reason is that the value of a claim on the riskier bank will be worth less in terms of consumption at a distant point. Consequently, consumers will send the notes of the higher-risk bank back for redemption. Thus an important conclusion is that higher-risk banks at a given location are monitored via more frequent note redemptions. The redeemability of notes means that bank type (asset risk) can be checked very quickly. This monitoring mechanism supports the equilibrium in which all seasoned banks (at a given location) have the same risk.

Banks the same distance away will have notes trading at the same discount. A bank with notes trading at a higher discount is either a seasoned bank that became insolvent or a new bank that must adjust its balance sheet to reduce its risk to be consistent with the risk taken on by its seasoned cohort. When the information about the ability of a new bank to honor notes is transmitted to distant locations, the price of its notes should adjust, contributing to the formation of the bank's reputation. The argument, thus, addresses an apparent paradox in free bank note prices, namely, that all solvent, seasoned, banks at a given location have notes trading at the same price. This is a result of the fact that bank notes functioned as a medium of exchange.

Section 3.4 empirically examines the predictions of the argument above as a prelude to using the prices of seasoned cohorts as a benchmark for the subsequent analysis. In particular, I examine whether the bank notes of seasoned solvent banks at particular locations, in fact, trade at the same price. I also look for evidence that higher-risk new banks' notes tend to be sent for redemption.

In Section 3.5 the main hypothesis of interest is tested, namely, the question of whether the notes of new banks are discounted more heavily than the notes of seasoned peer banks. In addition, I investigate whether the prices of new banks' notes are fair lemons premia. The size of the initial discount on new banks' notes relative to the discount on the notes of seasoned peers, the lemons premium, depends on the degree of adverse selection. If cross-section variation in public and private banking arrangements in different states affects the degree of adverse

selection, then this should be reflected in the initial discounts on new banks' notes. This is examined in Section 3.6. Section 3.7 examines whether the initial note prices differentiate between banks that subsequently go bankrupt and those that do not. In other words, whether there is evidence of a reputation effect or not, market participants may have sufficient information to distinguish between banks of different types. Section 3.8 examines the issue of technological change. The introduction and spread of the railroad and the telegraph may alter the ability of market participants to monitor banks and price notes. An index of technological progress is introduced and used to analyze the effects of technological change on the ability of market participants to discipline banks. Section 3.9 offers a conclusion.

3.2. PRE–CIVIL WAR BANK NOTE MARKETS

In pre–Civil War America, banks could open by obtaining a charter from a state legislature and satisfying state regulations concerning capital and reserves or, if the state allowed free banking, by depositing specified (state) bonds with a state regulatory authority, allowing them to issue private money.[7] If a free banking law was passed, then free and chartered banks could coexist if free banks entered the industry. During the Free Banking Era, 18 states adopted a version of free banking and 15 retained the chartered banking system.

All banks (free and chartered) issued distinct private monies, bank notes. Notes were issued in convenient denominations to facilitate use as media of exchange. Bank notes were pervasively used as a medium of exchange because there was no viable alternative medium. For example, Gouge (1833, p. 57) wrote that "of large payments, 999 in a 1,000 are made with paper. Of small payments, 99 in a 100. The currency of the country is . . . essentially a paper currency." With a well-functioning government currency system, bank notes might be dominated, but during the antebellum period, the costs of using specie were sizable. The government did not print paper money, and there were problems with the available coins. Not only was specie difficult to transport, but many coins were foreign, so there was a confusing array of denominations. There was no domestic coin between the 50-cent piece and the $2.50 gold dollar. Moreover, the law did not provide for the reminting of underweight coins,

7. "Free banking" refers to the passage of a general incorporation law for commercial banks. Free banking laws varied by state but tended to incorporate some common features. Typically, banks had to back their note issuance with designated state bonds deposited with state regulatory authorities. Also, bank notes were printed and registered under the direction of the regulatory authorities. Further background can be found in Cleaveland (1857), Grant (1857), Dewey (1910), and Hammond (1957).

which meant that coins might have a negative rate of return (see Carothers 1930).

Banks issued notes to finance loans, mortgages, and security purchases (mostly state bonds). The notes then circulated as media of exchange. At a bank's home location, the notes circulated at par because of the redemption option; at the home location of the issuing bank, any note price below par would result in the immediate exercise of the option allowing the note holder to obtain specie (if the bank was solvent). Consequently, all transactions using the notes of banks at that location would be conducted at par, consistent with Fama (1983), who argued that this would be the case for non-interest-bearing private monies.

It is not clear whether bank notes circulated across different states and regions in significant amounts. Unfortunately, there is no direct evidence in the form of note volumes that can be brought to bear on this question. The qualitative evidence, however, is highly suggestive. First, during this period, there were large interregional trade flows.[8] Some of this trade appears to have been conducted with bank notes because of the transportation costs of using specie (see the discussion in Atherton [1971]). The literature of the time repeatedly makes this point. For example, "Bank paper is 'convertible' into silver only, which is inconvenient for large payments, and for transportation to distant places in large amounts" (Gouge 1833, p. 59). There are many examples in which the observer reports the common use of distant notes to conduct trade. For example, in 1864 one observer commented that "there are no less than one thousand different kinds of bank notes, which every businessman in New York or New England is called upon to criticize and examine, and pay discount on, and suffer more or less, in the ordinary course of trade" (Shepard 1864). Or, in another case, "In April, 1838, the circulation of the northern portion of Wisconsin Territory was made up almost wholly of the notes of the banks organized under the general banking law of Michigan" (Merritt 1900). Green (1972) makes the point that Louisiana banks' notes circulated widely throughout the South. See also Atherton (1971).

Such observations are consistent with the fact that newspapers reporting the prices of bank notes, called "bank note reporters," were published in all major cities and were also consulted in rural areas (see Dillistin 1949). Bank note reporters were exhaustive in their coverage; that is, they reported a price for every existing private money in North America. The bulk of such newspapers was devoted to listing these prices together with descriptions of counterfeits.

8. Interregional trade flows in antebellum America were sizable (see Pred 1980; Mercer 1982). Fishlow (1964) presents quantitative evidence on the size of these flows, and Lindstrom (1975) specifically discusses Philadelphia.

Demand for these newspapers is consistent with notes' traveling some distance in the course of trade.

3.2.1. Bank Note Price Data

Note prices represented a system of fixed exchange rates with wide bands. Notes were redeemable in specie (at par), but only at the location of the issuing bank. For transactions at a distance away from the issuing bank, the price of a note could be below par since arbitrage via the redemption option was costly because of the time it took to return to the issuing bank. Thus note prices of distant banks were quoted at discounts. These discounts reflected the risk of the bank's asset portfolio, leverage of the bank, and the time involved to take the note back to the issuing bank (see Gorton 1993).

Note prices or discounts were established in informal secondary markets, where note brokers traded notes. Note prices in the secondary market were reported by the bank note reporters, which were consulted when unfamiliar notes were used in a transaction or sold in the secondary market. Bank note reporters were competitive, with several sometimes operating in larger cities (see Dillistin 1949). The data used in this study are taken from *Van Court's Counterfeit Detector and Bank Note List*, a bank note reporter printed monthly in Philadelphia from February 1839 through December 1858.[9] *Van Court* was a small tabloid providing general business news together with the discounts from par on the notes of the banks of 29 states and territories and three provinces of Canada. In all, note prices of approximately 3,000 banks are provided. (Appendix table 3.B1 shows the coverage provided by *Van Court*.)

The prices reported by *Van Court* are in the form of discounts from par; that is, the number "3" means that a $1.00 note of that bank is trading for 97 cents worth of gold (see Gorton 1989*b*).[10] The prices are not necessarily transactions prices, and the volumes traded are not known. Nevertheless, it seems reasonable to believe that they are fairly accurate since it is known that merchants relied on such reporters and that the bank note reporter market was competitive.

The prices in *Van Court* refer exclusively to the Philadelphia secondary note market. At a different location, say Chicago, prices would differ (even for a

9. See Gorton (1989*b*) for a more detailed description of *Van Court's Counterfeit Detector and Bank Note List.*

10. All note denominations of a given bank were discounted from face value by the same amount, and there were no "volume" discounts.

bank with the same asset risk and leverage), as we shall see below, because the distances back to the issuing banks would differ.

3.2.2. Cross-Section Variation in State Banking Systems

The banking systems in the various states and territories differed in a number of important dimensions. Some states allowed entry into banking under free banking laws and some maintained exclusively chartered systems; some allowed branching; some provided insurance for circulating bank liabilities; and some had private arrangements among banks that were important.

A traditional hypothesis is that banking systems that passed free banking laws experienced more bank failures and larger losses than chartered banking systems did. Rockoff (1971, 1974, 1975), while stressing the heterogeneity of free banking experiences, finds some support for this view. Rolnick and Weber (1982, 1983, 1984) find little evidence of pervasive wildcat banking, arguing that falling asset prices are a better explanation of failures in free banking states. Rockoff and Rolnick and Weber do not directly compare the experiences of free and chartered systems, however. Kahn (1985) compares the experiences of four free banking states with two chartered systems and with New Jersey, which passed a free banking law midway through the period. He finds that free banking legislation "often resulted in very high failure rates in those states relative to failure rates in non-free-bank states" (p. 885), though Kahn stresses that this is based on ex post data.

It is important to emphasize that chartered banking states also had a variety of experiences. In particular, passage of free banking laws was not necessary for the rapid growth of banks. Kahn (1985) cites Maine and Maryland as examples. Other chartered states restricted entry; Rockoff (1974) cites Pennsylvania, Tennessee, and Missouri as examples.

Together the evidence of Rockoff and Rolnick and Weber strongly suggests that the earlier view that free banking was synonymous with wildcat banking is incorrect, but it remains less clear how free banking systems performed relative to chartered systems.

It is important to note that, besides differing as to whether free banking was allowed or not, state banking systems significantly varied in other ways as well. These other factors will subsequently be important in assessing whether initial note discounts priced the degree of adverse selection across different states. These other factors fall into two categories. First, some states allowed banks opportunities that seem to have raised their expected returns for the same risk. In particular, some states (Virginia, North Carolina, South Carolina, Georgia, and Tennessee) allowed branching, which made these systems less risky (see Schweikart 1987; Calomiris and Schwiekart 1988; Calomiris 1989). Also, some

states had successful state insurance systems (Indiana, Iowa, and Ohio), whereas others had less successful systems (New York, Vermont, and Michigan) (see Calomiris 1989).

A second factor concerns private bank monitoring arrangements. Banks in New England were part of the Suffolk System, a private coalition of banks centered around the Suffolk Bank of Boston, generally viewed as a quasi central bank. New England banks were apparently less risky because of regulation of their activities by the Suffolk Bank (see Whitney 1878; Dewey 1910; Mullineaux 1987).

Variation in characteristics of state banking systems suggests that the degree of adverse selection of new banks may vary, affecting the price of new banks' notes. Stricter entry requirements, whether formal (e.g., different capital and reserve requirements) or informal (as with the Suffolk System), might well have prevented "bad" banks from entering.

3.2.3. Defining "New" Banks

This study focuses attention on new banks issuing notes for the first time. As there is no other extensive information available, a "new" bank must be defined using *Van Court*'s published prices. In order to be useful to consumers, a bank note reporter such as *Van Court* had to have exhaustive coverage. Every conceivable note that might be offered as payment in a transaction had to have a quoted discount or price. It is worth stressing that the bank note reporter market was competitive (see Dillistin 1949). Thus it seems reasonable to take the initial discount reported by *Van Court* on a bank's note as essentially the primary issuance price in Philadelphia. A new bank is defined, for purposes of this study, to be a bank appearing for the first time in *Van Court* after the first six months of publication.[11]

The definition of a new bank results in a sample of 1,673 banks that entered during the period. Figure 3.1 presents a bar graph of the number of new banks entering each year during the sample period. Entrants are, to some extent,

11. The first six months of publication are excluded because Van *Court*'s first issues were not apparently exhaustive in covering the existing banks. Initially, Van *Court* appears to have been expanding coverage to include banks that were seasoned but had not been included previously. The prices of many banks are listed in the first six months at the modal discount for that location, suggesting that they are not new. Including the first six months shows large numbers of banks as "new" compared to subsequent numbers of entering banks. Excluding the first six months eliminates 713 banks that would otherwise have been classified as new. That the remaining banks are, in fact, new was checked for a small sample of New York banks by comparing the state regulatory listings for banks not previously listed with *Van Court*'s new entries. This confirms that the banks are, in fact, new.

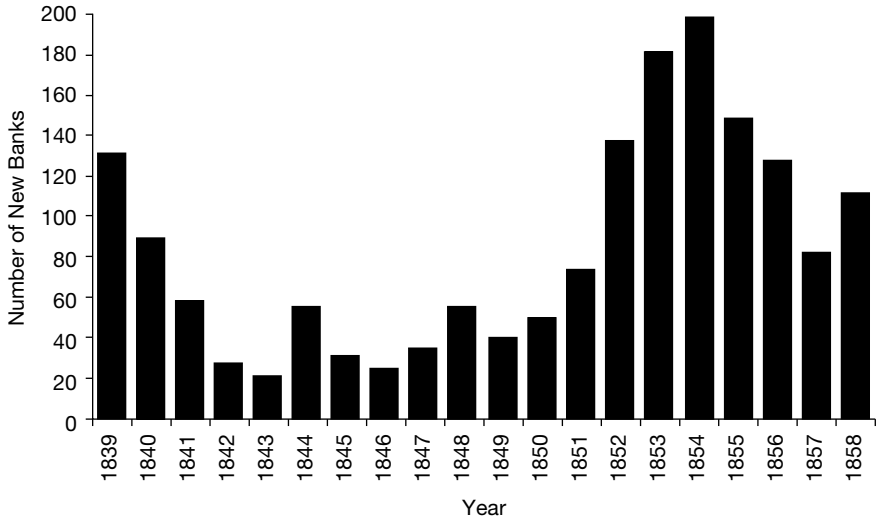


Figure 3.1 Number of new banks

clumped in the early period, when some states followed the lead of New York in adopting free banking, and in the early 1850s, when a number of additional states adopted free banking.

3.3. BANK NOTE PRICES, ARBITRAGE, AND MONITORING VIA REDEMPTION

A crucial step for the subsequent analysis is the proposition that *all seasoned banks at the same location will have identical note discounts at given distant locations, corresponding to identical asset risk (assuming capital requirements are binding so banks have the same leverage)*. In this section, I discuss this proposition informally; it is established formally in Appendix A in the context of a specific model. The proposition is stronger than the statement that note prices (discounts) must accurately reflect the default risk of the issuing bank. While this is true, the proposition says that the asset risk of banks at a given location must be the same. I show that the reason is that notes are used as media of exchange. The mechanism that enforces the equilibrium is redemption of notes of banks that choose higher levels of risk than that of seasoned peers. This monitoring feature of note redemptions is induced by arbitrage possibilities that arise if a bank chooses higher asset risk than its peers. The proposition will be examined empirically in the next section and will become the basis for using the note prices of seasoned banks as benchmarks against which the notes of new banks can be compared. I shall also examine evidence of monitoring via redemptions in the next section.

3.3.1. Bank Note Discounts

Imagine an economy in which agents are spatially separated and trade because they have a preference for goods from other, more distant, locations. I assume that (i) bank notes are used as media of exchange; that is, they are used to satisfy a cash-in-advance constraint; (ii) bank notes are risky because they finance risky assets; the issuing bank may fail to honor its notes at par if they are presented for redemption; and (iii) the further a bank's note is away from the issuing bank's location, the longer time it takes to return the note for redemption. The assumption that bank notes are used as media of exchange, assumption i, presumes that they are not dominated by another medium, such as specie. As discussed above, there was no government paper currency during this period, and trade with specie was costly. The assumption that banks are risky, assumption ii, should be interpreted further to mean that the investment opportunity set of banks and the cost of capital are taken as given (i.e., it is optimal for seasoned banks to be risky). Assumption iii will be interpreted to mean that distance away from the issuing bank is equivalent to the time it takes to receive the risky payoff of a note redemption.[12] In other words, think of distance as the maturity of the risky note. With these assumptions we can ask how a bank's note price (discount from par) is determined at any given location.

It is easy to price the note of a bank when the note is at the same location as the issuing bank. At the location of the issuing bank, its notes must trade at par because, if not, there is an arbitrage opportunity since it is costless to redeem the note at the bank (the time it takes to return to the issuing bank is zero). But if a particular bank's note moves further away from the bank's location in the course of trade, then a discount from face value will arise along the way (it is this capital loss that would make notes dominated if there were a superior alternative). The reason is that, from distant locations, it takes time to return the note to the location of the issuing bank, and the bank is risky. Pricing the note in this context is equivalent to pricing a risky pure discount bond in which the maturity is equal to the time it takes to return to the issuing bank. In fact, at first glance, it would seem that the notes of different banks at the same location could be priced differently at some particular distant location (i.e., maturity), as long as the different prices reflected the different default risks. This would be true in efficient markets if notes were not used as a medium of exchange.

Now consider the implications of using notes as a medium of exchange. At any date a particular bank's note may be held by an agent to satisfy the cash-in-advance constraint or it may be sent back to the issuing bank for

12. If the issuing bank is a distance d away, then assume that the maturity of the note is d periods, ignoring, for simplicity, the fact that there is a round-trip.

redemption (i.e., the agent will receive a risky payoff some periods from now, depending on how far away the issuing bank is located). If the agent is indifferent between these two alternatives, then the note may again be priced as a risky (pure discount) debt claim with maturity corresponding to distance away from the issuing bank. Otherwise a price bound is established. In Appendix A, conditions are provided under which a closed-form solution for note prices based on Black and Scholes (1973) can be derived. This pricing formula is useful because it shows that (as usual with bonds) the value of a note varies inversely with time to maturity, risk, and leverage.

The basis of the proposition is the fact that the value of a note declines as it moves further away from the issuing bank (because it then will take more time for the note to be returned for redemption). More specifically, a standard result on risky debt from contingent claims (see Merton 1974) is that the riskier the note (bond), the greater the decline in value as it moves further away (i.e., as the maturity increases). Since notes finance consumption purchases that may be made at locations further away from the issuing bank's location, the consumer will not be indifferent between the notes of two banks of different risk an equal distance away. If the consumer moves still further away from the issuing banks' location, increasing the time to redemption (maturity), the riskier banks' notes will decline in value by relatively more; hence the consumer exchanges fewer consumption units when shopping at the distant location. A less risky bank's notes will be preferred as a medium of exchange and the riskier bank's notes will be sent for redemption. But then equilibrium requires that the notes of all banks at a given location have the same risk, and none are sent for redemption. If banks could produce riskless liabilities and still earn the required rate of return on bank equity, then such notes would predominate. Of course, if using specie is less costly, then it might dominate notes. The proposition describes a world in which these alternatives are not available.

3.3.2. Discounts and Monitoring

Establishment of the equilibrium in which all banks at a given location have notes trading at the same prices relies on the argument that the notes of a higher-risk bank, at a given location, will be redeemed. Because a riskier bank will face more redemptions, it would have to hold more reserves or become insolvent. Since reserves are not interest-bearing, a bank with more reserves would be less profitable. Thus any difference in note prices induces a natural monitoring mechanism, namely, note redemptions. The mechanism of redemptions establishes the equilibrium quality (risk) of banks, resulting in the circulation of seasoned banks' notes at the same price without redemption.

Privately each bank may have an incentive to increase risk (above the equilibrium level of risk of bank portfolios at its location), that is, to be a "wildcat bank." Increasing risk will increase the value of the bank's equity, but market participants, recognizing the incentives of the bank, will discount its notes appropriately, penalizing the bank when it first introduces the notes into the market (the lemons premium in Diamond's [1989] model). Since the new, possibly wildcat, bank chooses a level of risk higher than the seasoned banks at its location, its notes will have a higher discount. In that case, by the argument above, all its notes will be redeemed. Redemption results in verification of bank type by establishing the ability of the bank to honor its notes with reserves, borrowings from other banks, or asset sales to other banks. If redemption occurs fast enough, wildcatting will not be profitable. The threat of redemption can prevent wildcat banking. Redemption corresponds to monitoring in Diamond (1991). This argument is formalized in Appendix A.

In the context of the Diamond (1989) setting, the arguments above should be interpreted as follows. The notes of new banks, to the extent that they are perceived to be riskier than seasoned peers, will be returned more frequently; that is, they will not circulate to the same extent. Redemptions serve the purpose of monitoring the new banks since if they are not good types, they will become insolvent faster. Thus, while new banks' notes will have higher discounts initially compared to those of seasoned peers, over time good banks and bad banks can be separated, and the type that can choose between a risky and a safe project will have an incentive to choose the low-risk project.

3.4. THE ENFORCEMENT OF ONE DISCOUNT PER LOCATION: EMPIRICAL EVIDENCE

The proposition says that the notes of banks at a given location will trade at the same price because, if they do not, the riskier banks will face redemptions until they adjust their asset risk or go bankrupt. In this section these predictions are examined empirically as a prelude to testing for the presence of reputation formation.

3.4.1. Do Seasoned Solvent Banks Face the Same Discount?

To examine the prediction that seasoned solvent banks' notes (at a given location) trade at the same discount, table 3.1 provides the average of the monthly percentages of total banks, at representative selected locations, whose notes were trading at the modal discount for each year.[13] The states shown in table 3.1

13. Gorton (1989*b*) contains the full set of results.

Table 3-1. PERCENTAGE OF BANKS WITH NOTES AT THE MODAL DISCOUNT: SELECTED STATES

	Connecticut		Georgia		Louisiana		Massachusetts	
	Modal Percentage	**Number of Banks**	**Modal Percentage**	**Number of Banks**	**Modal Percentage**	**Number of Banks**	**Modal Percentage**	**Number of Banks**
1839	84.03	42	63.69	25	95.26	20	98.49	136
1840	97.35	42	57.81	26	95.16	21	100.00	135
1841	96.73	42	54.19	18	96.06	19	100.00	130
1842	94.42	41	77.95	20	52.10	20	97.88	133
1843	95.00	40	63.40	18	50.88	20	96.32	133
1844	98.37	42	87.33	19	47.42	21	97.03	132
1845	98.16	42	85.28	28	50.00	20	97.74	133
1846	98.75	40	86.67	20	52.63	19	97.44	133
1847	99.58	40	89.76	18	52.63	19	98.80	110
1848	100.00	37	78.89	14	50.00	18	98.80	112
1849	100.00	40	83.98	13	79.66	18	99.54	122
1850	100.00	44	94.87	13	100.00	8	100.00	129
1851	97.94	47	77.57	13	100.00	8	100.00	133
1852	99.36	56	96.80	14	100.00	6	99.92	141
1853	99.42	63	96.77	18	100.00	8	100.00	150
1854	99.48	69	82.01	16	100.00	10	100.00	156
1855	100.00	69	97.02	18	100.00	10	100.00	162
1856	100.00	73	60.63	25	100.00	9	100.00	164
1857	96.27	77	64.84	24	100.00	8	99.75	175
1858	87.87	81	58.97	30	100.00	11	99.27	179

	New York City		New York State*		Ohio		Philadelphia	
	Modal Percentage	Number of Banks	Modal Percentage	Number of Banks	Modal Percentage	Number of Banks	Modal Percentage	Number of Banks
1839	93.71	41	78.33	148	89.89	38	100.00	42
1840	94.71	43	92.07	181	83.51	42	96.06	39
1841	85.43	38	68.13	168	84.13	40	82.92	39
1842	80.78	41	78.25	164	71.71	34	59.17	32
1843	73.51	39	67.50	166	67.72	36	70.83	30
1844	83.49	39	82.63	183	61.38	35	78.00	29
1845	89.15	36	83.35	184	70.48	35	94.09	26
1846	80.09	36	77.58	185	80.07	40	94.12	25
1847	78.70	36	76.89	203	81.52	39	93.44	22
1848	84.25	34	76.84	212	82.41	44	93.33	22
1849	100.00	29	81.30	209	77.27	44	93.68	21
1850	99.18	32	86.79	209	76.28	44	93.75	21
1851	97.50	41	87.35	238	76.85	43	93.75	21
1852	97.43	49	96.21	234	92.86	30	93.75	21
1853	98.18	64	87.93	286	94.60	39	100.00	21
1854	97.68	68	95.85	309	100.00	37	100.00	20
1855	88.29	68	96.44	318	93.32	37	100.00	20
1856	92.23	70	96.57	337	91.36	38	100.00	20
1857	93.38	68	95.86	320	87.12	38	100.00	20
1858	98.28	58	84.11	283	81.46	36	94.52	20

NOTES: The modal percentage is the average of the 12 monthly modal percentages (percentage of total banks with notes trading at the modal discount). The number of banks is the number of banks in existence during the year.

* All banks in New York State excluding New York City banks.

are representative geographically and with respect to type of banking system. At each date the bank notes of most banks at each location are trading at the same discount in the Philadelphia note market, the modal discount. It is clear from the table that at most locations the percentage of banks with notes trading at the same discount in Philadelphia is extraordinarily high.

In almost every case, the notes of other banks, not trading at the modal discount, are trading at higher discounts, usually much higher, suggesting that these notes are claims on insolvent banks (see Gorton 1989*b*).[14] When a bank went bankrupt, state bank regulators liquidated the bank over a period of time, usually some years. During this time the bank's notes could continue to circulate, but they would be equity claims on the bank. Consequently, these notes would trade at "deep" discounts. To investigate this, table 3.2 provides the modal discounts, averaged over the months of each year, and the average nonmodal discount.[15] It can be seen that the nonmodal discounts are typically much larger than the modal discounts.[16] As expected, in Philadelphia, the modal discount is always zero, indicating that bank notes trade at par at the home location. Also, notably, even states such as New York, where free banks and chartered banks covered by state insurance coexisted, the discount on the notes of all solvent banks is the same!

The high percentages of banks with notes trading at the modal discount are consistent with the proposition above. Banks not trading at the modal discount are insolvent.

3.4.2. Evidence of Monitoring

The argument above also predicts that the notes of a new bank that are trading at a discount higher than the modal discount of seasoned peers at their location will be redeemed more frequently. In the face of such redemptions, we would expect "bad" banks, that is, high-risk banks, to be detected fairly fast. In fact, the notes of banks of higher perceived risk would not circulate as far.[17]

14. This was verified for a small sample of New York State banks.

15. The reader will note some negative entries in table 3.2. They occurred during periods of suspension of convertibility (during the banking panics of 1839 and 1857). During a period of suspension, it was not possible to obtain gold in exchange for notes. *Van Court* essentially changed the numeraire from gold to Philadelphia bank notes during these periods. Thus a negative number indicates a *premium* in terms of Philadelphia banks' notes. See Gorton (1989*b*) for a more complete discussion.

16. In a few cases, such as Connecticut in 1851 and Georgia in 1850, a single bank's notes traded at a discount lower than the modal discount for a few months. In no case is the nonmodal discount systematically lower than the modal discount.

17. In terms of the model in Appendix A, with a higher σ, the optimal d that solves (3A.4) is lower.

Table 3-2. MODAL AND NONMODAL DISCOUNTS: SELECTED STATES

	Connecticut		Georgia		Louisiana		Massachusetts	
	Modal Discount	Nonmodal Discount	Modal Discount	Nonmodal Discount	Modal Discount	Nonmodal Discount	Modal Discount	Nonmodal Discount
1839	.045	−.021	5.975	7.754	3.773	13.125	−.318	.313
1840	−3.083	4.615	7.375	11.035	2.417	25.000	−3.083	10.833
1841	−1.500	8.906	8.917	16.516	4.125	25.000	−1.917	. . .
1842	−.167	19.315	9.167	13.308	18.337	27.979	−.167	56.515
1843	.833	21.708	3.750	10.333	2.542	50.827	.833	54.846
1844	.500	3.500	2.000	14.286	1.500	41.302	.500	55.417
1843	.500	5.000	2.000	13.667	2.000	44.667	.500	27.692
1846	.500	5.000	1.833	15.548	2.500	38.333	.500	34.194
1847	.500	5.000	1.229	16 818	1.250	38.333	.500	60.000
1848	.500	. . .*	1.833	3.818	1.083	38.333	.500	60.000
1849	.430	. . .	1.375	1.900	1.833	35.930	.430	43.000
1850	.380	. . .	1.000	.750	1.438	. . .	.380	. . .
1851	.380	.250	.979	1.036	1.104	. . .	.380	. . .
1852	.326	.500	1.021	1.250	1.229	. . .	.326	.380
1853	250	38.750	.885	2.750	.917	. . .	.250	. . .
1854	.388	3.000	1.063	3.711	1.021	. . .	388	. . .
1855	.313	. . .	1.208	1.250	1.792	. . .	.313	79.500
1856	.250	. . .	1.000	2.000	1.917	. . .	.250	53.347
1857	.229	17.827	2.042	3.624	1.021	. . .	.229	5.607
1858	.295	8.623	1.542	7.780	1.313	. . .	.295	3.167

Table 3-2. CONTINUED

	New York City		New York State†		Ohio		Philadelphia	
	Modal Discount	Nonmodal Discount	Modal Discount	Nonmodal Discount	Modal Discount	Nonmodal Discount	Modal Discount	Nonmodal Discount
1839	−2.0450	−.026	−.795	−1.276	4.409	2.368	.000	. . .
1840	−4.0330	31.239	−2.375	1.770	4.833	4.407	.000	21.667
1841	1.0630	30.356	−1.583	16.694	7.250	8.706	.000	24.927
1842	−.8750	31.500	.292	32.623	10.167	23.556	.000	38.561
1843	.1670	29.600	.197	42.599	2.792	38.727	.000	33.214
1844	.0210	22.071	.750	40.290	1.646	24 600	.000	22.058
1845	.0000	28.813	.750	33.893	2.021	25.565	.000	28.636
1846	.0000	47.244	.813	26.813	2.125	30.025	.000	24.167
1847	.0000	51.914	.750	24.667	1.333	38.838	.000	18.917
1848	.1670	50.159	1.010	24.376	2.083	42.418	.000	14.833
1849	.1250	. . .	.802	19.784	1.625	48.000	.000	12.333
1850	.0100	.750	.750	9.350	1.448	49.500	.000	13.833
1851	.0000	.750	.750	9.811	1.271	49.328	.000	15.000
1852	.1250	.589	.688	12.495	1.271	70.000	.000	15.000
1853	.1250	.500	.510	8.742	.979	31.155	.000	. . .
1854	.1250	.097	.542	12.197	1.866	. . .	.000	. . .
1855	.1560	3.643	.542	14.813	1.475	35.397	.000	. . .
1856	.1250	4.903	.500	11.725	1.000	20.000	.000	. . .
1857	.0104	40.271	.458	16.891	2.250	30.198	.000	. . .
1858	.8960	15.000	.375	6.638	1.208	29.938	.000	60.000

NOTES: The modal discount is the annual average of the 12 monthly modal discounts. Similarly, the nonmodal discount is the average of the monthly nonmodal discounts.

* Indicates that all the banks during the months of that year had notes trading at the modal discounts.

† All banks in New York State excluding New York City banks.

Consequently, learning by market participants should happen fairly fast. Moreover, as a consequence of redemptions, all new banks should hold more reserves in anticipation of redemptions, a prediction examined in a subsequent section. While there are no data available on redemptions at individual banks, some evidence that this is the case can be adduced by examining how fast bad new banks are detected.

In order to examine this issue, the sample of new banks must be split into good and bad banks. To define a "good" bank I shall rely on the prediction that such a bank eventually has notes priced the same as those of seasoned peers. Therefore, a good bank is defined to be a bank whose note discount has converged to the modal discount (at that location) 13 months after entry.[18] Other banks are deemed bad banks (their note discounts become increasingly larger than the modal discount as time goes by).

A bank that becomes insolvent is treated by *Van Court* in one of two ways. Either its notes continue to trade at high discounts, since they are essentially equity claims at that point, or the discount on the bank's notes is no longer reported. A bank whose discount is initially in excess of the modal discount may eventually (after 1 year by the definition above) become a good bank. Suppose it is assumed that new banks that *Van Court* drops from newspaper coverage are bad banks that have been detected. In fact, just prior to being dropped, these new banks have higher discounts than other new banks, suggesting that they did become insolvent. The percentage of new bad banks that *Van Court* discontinues reporting on provides a *lower bound* on the number of bad banks that have been detected.

Examining the percentage of new bad banks that *Van Court* discontinues reporting on provides some sense of the speed with which bad banks are detected. Table 3.3 presents some (representative) such evidence. As can be seen in the table, for many states, over 50 percent of the bad banks are detected within the first year of their existence. The states in which no bad banks have been detected within the first year are states that are distant from Philadelphia and have few banks, Alabama and Nebraska. In the case of Delaware, there is only one bad bank. On the other hand, bad banks are detected very fast in states with large numbers of bad banks, such as New York and Indiana. The evidence in table 3.3 is consistent with the notion that bad banks are forced into insolvency via redemptions since their assets typically would have maturities longer than 1 year.[19]

18. The requirement is that the discount of the entrant be at the mode for three consecutive months, 13, 14, and 15 months after entry. Results are not particularly sensitive to a variety of other definitions of good and bad banks.

19. See Dewey (1910) for a discussion of the loans made by banks during this period.

Table 3-3. SPEED OF DETECTION OF BAD BANKS

State	Number of Bad Banks	Percentage of Bad Banks Surviving after: 2 Months	3 Months	6 Months	12 Months	16 Months
Alabama	4	100.00	100.00	100.00	100.00	100.00
Connecticut	12	100.00	100.00	100.00	66.67	16.67
Delaware	1	100.00	100 00	100.00	100.00	100.00
District of Columbia	21	57.14	52.38	42.86	19.05	9.52
Georgia	47	93.62	82.98	70.21	59.57	31.92
Illinois	30	93.33	93.33	73.33	36.67	26.67
Indiana	113	70.80	69.90	65.50	53.10	47.80
Louisiana	12	100.00	91.67	83.33	41.67	41.67
Maine	59	52.54	52.54	42.37	37.29	30.15
Maryland	24	100.00	95.83	87.50	87.50	79.17
Massachusetts	19	100.00	89.47	84.21	57.90	21.05
Michigan	46	45.65	45.65	41.30	30.44	23.91
Mississippi	19	36.84	31.58	31.58	31.58	31.58
Montana	11	100.00	90.91	81.82	54.55	9.09
Nebraska	9	100.00	100.00	100.00	100.00	100.00
New Hampshire	20	95.00	95.00	65.00	15.00	10.00
New Jersey	25	92.00	84.00	76.00	56.00	40.00
New York	256	67.58	63.67	60.55	56.64	47.27
North Carolina	0	...	...	...	...	...
Ohio	15	93.33	80.00	66.67	46.67	33.33
Pennsylvania	54	96.30	90.74	70.37	53.70	37.04
Rhode Island	7	100.00	100.00	100.00	57.14	14.29
South Carolina	7	100.00	100.00	100.00	85.71	71.43
Tennessee	36	97.22	91.67	83.33	75.00	63.89
Vermont	4	100.00	100.00	100.00	75.00	50.00
Virginia	39	100.00	100.00	94.87	87.18	79.49

I shall now turn to testing the main prediction of the reputation theory, that banks issuing notes for the first time should face higher discounts on their notes than banks at the same location that have been in existence for some time.

3.5. REPUTATION FORMATION AND THE PRIMARY NOTE MARKET

We are now in a position to ask whether the notes of new banks are discounted more heavily than those of seasoned peer banks at that location. We shall examine the discounts on new banks' notes compared to the modal discount of

banks at that location. The "excess entry discount" for new banks, entering the market at time *t,* at a particular location, is defined to be

$$\frac{\text{entry discount}_t - \text{modal discount}_t}{|100 - \text{modal discount}_t|}.$$

The excess entry discount is the difference between the discount on the notes of a new bank, entering at time *t,* and the modal discount for seasoned banks at that location at time *t,* normalized for the price of the seasoned banks' notes at time *t* (to facilitate comparison across time and location).

The advantage of this definition of the excess entry discount is that many observed factors are indirectly accounted for by their influence on the modal discount. For example, if a state changes its bank regulations, if it introduces free banking, or if there is a macroeconomic shock, the modal discount will change. Gorton (1993) argues that the modal discounts are accurate reflections of such risk factors. Thus the benchmark is quite robust.

3.5.1. Discounts on the Notes of New Banks

The main prediction of Diamond's reputation theory is that the excess entry discounts should be significantly positive because the notes of new banks must offer a premium to note holders (over the rate offered on the notes of seasoned banks) to induce them to hold them since new banks do not have credit histories. To examine this question the excess entry discount is computed for all new banks during the period; there are 1,673 new banks. A finding of a significantly positive excess entry discount would be evidence in favor of the theory.

Panel A of table 3.4 provides the average excess entry discount on the notes of all new banks that entered during the period. Also provided are the results of the test that the excess entry discount is significantly different from zero. As can be seen, the average excess entry discount is significantly positive as predicted by the reputation model of Diamond. This is also true of subperiods, as shown in panel B of table 3.4.[20]

3.5.2. Are the Excess Entry Discounts Fair Lemons Premia?

If new banks are, in fact, riskier, on average, than seasoned banks and the higher discount accurately reflects this risk, then a market participant buying a portfolio

20. The subperiods are chosen to correspond to the measure of technological change introduced in Section 3.8.

Table 3-4. EXCESS ENTRY DISCOUNT

Period	Mean Excess Entry Discount	Number of New Banks	Standard Deviation	Minimum	Maximum	*t*-Statistic
			A. ALL BANKS			
1839–58	.0258	1,673	.110	−.286	1.290	9.56
			B. BY PERIOD			
1839–45	.0697	412	.171	−.059	1.290	8.26
1846–50	.0220	203	.107	−.021	.797	2.94
1851–58	.0080	1,058	.068	−.286	.737	3.96

of the notes of new banks at the date of entry should not earn a higher return compared to a portfolio of seasoned banks' notes purchased at the same dates and locations. That is, the discounts should be fair "lemons premia" since some of the new banks will fail and some will not. Thus a portfolio of new banks' notes should include some notes that suffer capital losses (when the bank fails or when information that it is a bad bank is revealed) and some notes that realize capital gains (when it is revealed to be a good bank).

To examine this question I form a portfolio of each new bank's notes at the date the new bank enters and examine the return on this portfolio over the first year of the bank's existence. The return on this portfolio is compared to the return on a benchmark portfolio composed of seasoned peer banks' notes as follows. On each date that a new bank enters, the benchmark portfolio purchases the note of a seasoned peer from that location. The benchmark portfolio is then held for a year. We can examine the difference in the returns on these portfolios. Thus, for a new bank entering at date *t,* the difference in returns is given by

$$\frac{P_{Nt+12} - P_{Nt}}{P_{Nt}} - \frac{P_{St+12} - P_{St}}{P_{St}} \equiv R_N - R_S,$$

where P_t is the price of the note at date t (100 minus the discount) and N and S refer to the new bank and the seasoned bank, respectively.

Table 3.5 reports the differences in returns between the two portfolios for the whole period and for subperiods. In each case the difference is insignificantly different from zero. The discounts on the notes of new banks appear to be fair since they provide the market rate of return on seasoned banks' notes. In this sense, there is no underpricing of new banks' notes.

3.5.3. Counterfeiting

The fact that the excess entry discounts are significantly positive, on average, and that they represent fair lemons premia does not, however, allow the immediate

Table 3-5. Return Differences between Portfolios of New Banks' Notes and Portfolios of Seasoned Peer Banks' Notes

	1839–58 (*N* = 1,673)	**1839–45 (*N* = 412)**	**1846–50 (*N* = 203)**	**1851–58 (*N* = 1,058)**
Mean return difference	−.0045	−.0046	.0023	−.0063
Standard deviation	.114	.122	.110	.105
Minimum difference	−1.045	−.983	−.443	−1.045
Maximum difference	2.240	2.240	1.000	.0301
t-statistic	−1.720	−1.090	.328	−1.68

conclusion that the lack of a credit history is the explanation. A non–mutually exclusive alternative hypothesis concerns counterfeiting of bank notes. Counterfeiting during the Free Banking Era was a serious problem (see Dillistin 1949; Glasner 1960). *Van Court* reports descriptions of counterfeit notes for every bank with a reported note price, suggesting that counterfeiting was widespread.

The result that the notes of new banks are more heavily discounted than the notes of seasoned banks at the same location is consistent with the interpretation that new notes were more likely to be counterfeits. It may have taken time for note holders to learn to recognize counterfeits of new notes. If the probability that a new bank's note is counterfeit was higher or if the public was less capable of recognizing counterfeits of new notes, then these notes would face higher discounts. As the public learns that the new notes are from legitimate banks and comes to recognize the counterfeits of new banks' notes, the excess entry discount would shrink. Learning about counterfeits is also tantamount to the acquisition of a reputation, but this reputation is conceptually distinct from the notion of a reputation proposed by Diamond.

There are several reasons why counterfeiting does not seem a persuasive explanation of the results in tables 3.4 and 3.5. First, a difficulty with the counterfeiting explanation of the results is that it is not clear that the notes of new banks would be more likely to be counterfeited than the notes of seasoned banks. There are costs to counterfeiting the notes of new banks. The main problem is that many of these banks become insolvent fairly quickly (as shown in table 3.3), making counterfeiting the notes of new banks very risky. Moreover, as we have seen in table 3.4, new banks' notes were more heavily discounted, making it less profitable to counterfeit them. Contemporaries of the period repeatedly observe that almost all notes were counterfeited, but that notes of "better" banks were more likely to be counterfeited. The *New York Times* observed in 1862 that

> out of the thirteen hundred and eighty-nine banks in the United States, only two hundred and fifty-three have escaped the attempts at imitation by

> one or another of the many species of frauds. And out of these two hundred and fifty-three, at least one hundred and forty-three are not worth counterfeiting, so that in round numbers, out of 1,300 bank note issues, but one hundred are not counterfeited. The rule is, that the better the bank, the more the counterfeits.
>
> [Quoted in GLASNER (1960, pp. 85–86)]

A second point concerns how counterfeiting was actually accomplished. The dominant method was not engraving, printing, photographing, or otherwise creating replicas of real notes. These technologies were expensive and not widespread. Instead, rather than the replication of notes, the predominant method involved the alteration of existing notes.[21] A typical method was to raise the denomination of an existing note, for example, by turning a $1.00 bill into a $10 bill by adding a zero. Another common method was to alter a note of an insolvent bank (trading at a high discount) so that it appeared to be a note of a solvent bank, thereby capturing the difference in the discounts. One observer writes as follows:

> There are now in circulation nearly four thousand counterfeit or fraudulent bills, descriptions of which are found in most Bank Note Lists. Of this number, a little over two hundred are engraved imitations—the residue being in point of general design entirely unlike the real issues of the banks whose names have been printed on them. These spurious notes—more properly altered—bills are generally notes of broken or exploded banks, which were originally engraved and printed by bank note engravers for institutions supposed to be regularly organized and solvent. [*Descriptive Register of Genuine Notes* (1859), cited by Glasner (1960, p. 82)]

Basically, the available counterfeiting technology, altering existing notes rather than printing new notes, restricted the choices of counterfeiters. It was not possible to focus counterfeiting activity exclusively on new notes. Attention was focused on notes that were poorly designed or poorly printed, which made alterations easier, or on notes that were more profitable to alter. Moreover, to the extent that activity could be focused, the available evidence suggests that it was the seasoned banks' notes that were more profitable to counterfeit. The conclusion is that counterfeiting cannot be the explanation for the results in tables 3.4 and 3.5. In fact, new banks' notes were *less* likely to be counterfeit.

21. Dillistin (1949) provides a discussion of the ways in which notes were altered and provides pictures of real and altered notes.

3.6. CROSS-SECTION VARIATION IN STATE INSTITUTIONS AND THE DEGREE OF ADVERSE SELECTION

Variation of excess entry discounts across states is likely to depend, in part, on the ability of banks to engage in risk taking. That is, the degree of adverse selection in an entering cohort may differ across states. As discussed above, the degree of adverse selection should depend on the public and private arrangements governing banking in the given state. This section examines these predictions.

3.6.1. Public and Private Banking Arrangements

Institutional factors that affect entry would be detectable in the excess entry discounts only if they affect the degree of adverse selection. It is important to keep in mind that these factors will also affect the benchmark of the modal discount if seasoned bank risk is affected (see Gorton 1993). So the excess entry discount will be affected only if these factors serve to deter bad banks from entering.

A state-run note insurance program may reduce the degree of adverse selection. New banks in states with successful state insurance programs should have lower excess entry discounts because these systems were mutual guarantee systems that included monitoring by other banks and state insurers (see Calomiris 1989). If monitoring by state regulators or by other banks is more intense in states with insurance programs, then fewer bad banks will enter the market. Calomiris divides these systems into successful insurance systems and unsuccessful insurance systems on the basis of their design and experience. In what follows I adopt his classification.

Also, as mentioned above, some states allowed branch banking, which evidence suggests reduced the bank failure rate, possibly because of diversification. The existence of branch banking would reduce the modal discount (a prediction confirmed by Gorton [1993]), but may also affect the excess entry discount. This would occur, for example, if competition from incumbents via branches raises the required quality of entrants in order to achieve success.

Private bank coalitions, in particular the Suffolk System of New England, should reduce the degree of adverse selection because participation in this system was a prerequisite for success. The Suffolk Bank, generally viewed as a quasi central bank, may have screened entrants. It appears that the Suffolk Bank was successful in reducing the risk of member banks. During the Panic of 1839 and its aftermath, only four out of 277 banks in New England (outside of Rhode Island) failed. In other areas of the country the failure rate was much higher. In Ohio, Illinois, and Michigan, 13.4 percent of banks failed.

The factors above would be important to the extent that they operated to reduce the proportion of bad banks in any entering cohort. Free banking laws, however, were designed to ease entry rather than restrict entry. Consequently,

the predictions about excess entry discounts with respect to whether the banking system is free or chartered are less clear. While a common conjecture is that since free banking made entry easier and that, consequently, the degree of adverse selection may well have been higher in free banking states, only Kahn (1985), who examined two chartered states, provides any evidence for this view, as discussed above.

When a free banking law was passed in a state, it did not necessarily mean that free banks entered. In every case in which free banks entered, they coexisted with chartered banks. In other words, there is no state in which chartered banks were forced out of the banking industry by competition from free banks. The argument above—that all note prices of banks at a given location will be the same—implies that when free banks enter under a new free banking law, either the new free banks' note prices will adjust to the price of the incumbent seasoned chartered banks or the opposite will occur. It cannot be the case, in equilibrium, that free banks and chartered banks coexist with notes trading at different prices. Indeed, in all states that passed free banking laws, solvent free and chartered banks traded at the modal discount for that location. A good example of this is New York, which had insured chartered banks and free banks coexisting for the entire period. (The free banks were not insured but faced bond backing requirements for note issuance.) Yet all these banks traded at the same discount when solvent.

Gorton (1993) found that the risk of banks (the asset value variance implied by the modal note price, found by inverting the Black-Scholes formula) trading at the modal discount was not affected by passage of a free banking law. This suggests that free banks and chartered banks coexisted because free banks adjusted their balance sheets so as to have the same risk as the incumbent chartered banks. It cannot be the case that seasoned chartered banks adjusted their risk levels to the anticipated level of risk that would prevail when free banks entered. By revealed preference, that level of risk could have been achieved without entry by free banks (if it could not have been achieved, then chartered banks would be driven out of the market, but this never occurred). One explanation for why free banks did not enter in some states that passed free banking laws might be that bank regulations prevented them from achieving the same risk level as the incumbent chartered banks. This is a question for further research.

While free banks that entered would have to adjust to the risk level of the incumbent chartered banks, the degree of adverse selection might be worse in free banking states. In that case the excess entry discounts would be larger because of the entry of more bad banks. In the four free banking states examined by Rolnick and Weber (1984), however, they do not find large numbers of banks failing in the first year. While it is not clear what "large" means since there is no benchmark for chartered banking states, it does not appear that there was a high proportion of wildcat banks entering. Rational wildcat bankers would not

enter in greater numbers if the threat of redemptions made it unprofitable (see Appendix A).

These observations suggest that the distinction between free and chartered banking systems may not help explain cross-section variation in excess entry discounts. Essentially, free banking laws while allowing entry may not necessarily result in the entry of large numbers of bad banks because of the threat of the redemption option when faced with competition from chartered banks.

3.6.2. Excess Entry Discounts and Institutional Factors: Tests

To examine whether the degree of adverse selection varies in the manner predicted, the excess entry discounts were regressed on the independent variables above, measured as dummy variables. If the banking system is a chartered banking system, the variable is set to one. If the state subsequently adopts free banking, then the chartered dummy variable is set to zero and the free banking dummy is set to one.

Table 3.6 presents the results of the regressions.[22] The cross-section variation of excess entry discounts by state does reflect risk factors that are expected a priori to play a role: branching, membership in the Suffolk System, and insurance reduce the excess entry discount. This is shown on the left-hand side of table 3.6, which presents a simple, time-series, cross-section regression of the excess entry discounts on new banks' notes on dummy variables for whether the state is a branching state, is a free or chartered banking state, has a successful or less successful insurance program, or is a state in the Suffolk System.

The regression includes two variables intended to capture business cycle variation: an index of stock prices and a dummy variable for suspension of convertibility.[23] Excess entry discounts are lower when the stock market goes up, possibly because new banks entered with more equity during these periods. The excess entry discount is not significantly affected by whether the new bank entered during a period of suspension of convertibility (suspension period). (The variable travel time is discussed below.)

With respect to whether the state allowed free banking or not, table 3.6 shows that there is no significant difference with respect to the degree of adverse selection. These dummy variables are significant for the period as a whole and for the early period (prior to 1846) but are not significantly different from each other. For the later periods, the variables are not significant. This is consistent

22. There are no intercepts in the regressions because all the dummy variables are used.

23. The monthly index of stock prices is taken from Smith and Cole (1935). A suspension period occurs during a banking panic, during which time all banks refuse to convert debt liabilities into specie on demand.

Table 3-6. CROSS-SECTION VARIATION IN EXCESS ENTRY DISCOUNTS

Independent Variable	1839–58 Parameter Estimate	1839–58 *t*-Value	1839–45 Parameter Estimate	1839–45 *t*-Value	1846–50 Parameter Estimate	1846–50 *t*-Value	1851–58 Parameter Estimate	1851–58 *t*-Value
	(1)		(2)		(3)		(4)	
Branching dummy	−.4100	−3.800	−.1340	−6.240	−.0780	−2.261	.0090	.746
Free dummy	.0610	3.500	.3480	8.200	.1300	.979	.0250	1.244
Chartered dummy	.0800	4.602	.3230	8.120	.1760	1.371	.0280	1.422
Good insurance	−.0300	−2.730	.0170	.630	−.0780	−2.293	.0030	.261
Bad insurance	−.0150	−1.920	−.1330	−7.400	−.0130	−.505	−.0002	−.027
Stock index	−.0004	−2.110	−.0010	−2.810	−.0010	−.748	−.0002	−1.101
Suffolk member	−.0290	−3.680	−.0530	−3.240	−.0770	−3.700	−.0050	−.596
Suspension period	−.0030	−.470	−.1060	−9.590	...	...	.0240	1.760
Travel time	.0003	6.580	...	...	...	...	...	...
R^2	.1032		.3113		.2009		.2224	
F-value	21.93		20.141		20.45		20.52	
Prob > *F*	.0001		.0001		.0001		.0001	
Degrees of freedom	1,637		410		194		1,033	

with the results of Rolnick and Weber (1984), who argued that free banking did not appear to have resulted in performance significantly different from that of chartered banking systems. The ex ante evidence from note market prices is in agreement with their ex post evidence concerning failures.

3.7. GOOD BANKS AND BAD BANKS

The result that the notes of new banks were, on average, discounted more heavily than the notes of seasoned peer banks provides evidence in favor of the reputation hypothesis. But it does not rule out the possibility that market participants could, at least to some extent, distinguish between "good" banks and "bad" banks. Perhaps there is enough prior information to allow such a distinction, even though there is not enough information to eliminate the significantly positive excess entry discount.

A good bank has been defined to be a bank whose note price eventually converges to the modal price (after 13 months by the definition above), whereas a bad bank is a bank whose note price diverges from the modal discount. Using this definition, we can ask whether the initial note discounts reflect the fact that the bank will subsequently turn out to be good or bad.

3.7.1. Market Distinctions between New Banks at Entry

To address the question of whether the market can distinguish between good and bad banks at entry, I separately compute excess entry discounts for good banks and bad banks (i.e., on the basis of their ex post performance). The question is whether the excess entry discounts are significantly different for the two groups. Table 3.7 shows the average excess entry discounts for all bad new banks entering during the period (col. 1) and all good new banks entering during the period (col. 2). Also shown are the computations for three subperiods. For the whole period as well as the subperiods, the excess entry discounts for the bad banks are significantly different from zero. For the good banks, the mean excess entry discount, while significantly different from zero for the whole period, is not significantly different from zero after 1845. During the later period (1846–58), entering good banks' notes are priced the same as (i.e., insignificantly different from) seasoned peers' notes.

The tests in panel B of the table show that for the whole period as well as subperiods, the mean excess entry discounts for the two groups are significantly different.[24] In other words, while the market significantly discounted

24. The tests in panel B of table 3.7 and in table 3.8 are tests of the equality of means, assuming that the samples are independent and have different population standard deviations (which is consistent with the different degrees of risk of bad banks and good banks). Consequently, instead of

Table 3-7. A. Excess Entry Discounts for Good Banks and Bad Banks

	All New				Subperiods			
	1839–58		1839–45		1846–50		1851–58	
	Bad Banks	Good Banks	Bad Banks	Good Banks	Bad Banks	Good Banks	Bad Banks	Good Banks
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mean excess entry discount	.0471	.0021	.124	.014	.086	.0005	.016	−.0004
Number of banks	881	792	178	133	51	152	552	507
Standard deviation	.147	.023	.200	.024	.200	.008	.091	.025
Minimum	−.286	−.286	−.011	−.011	−.015	−.021	−.286	−.286
Maximum	1.290	.211	.756	.167	.797	.091	.737	.211
t-value	9.490	2.560	8.270	6.500	3.090	.849	4.220	−.347

B. Tests of Difference of Mean Excess Entry Discount between Good and Bad Banks

	1839–58	1839–45	1846–50	1851–58
t'	8.96	7.27	3.05	4.07
Degrees of freedom	928	184	50	641

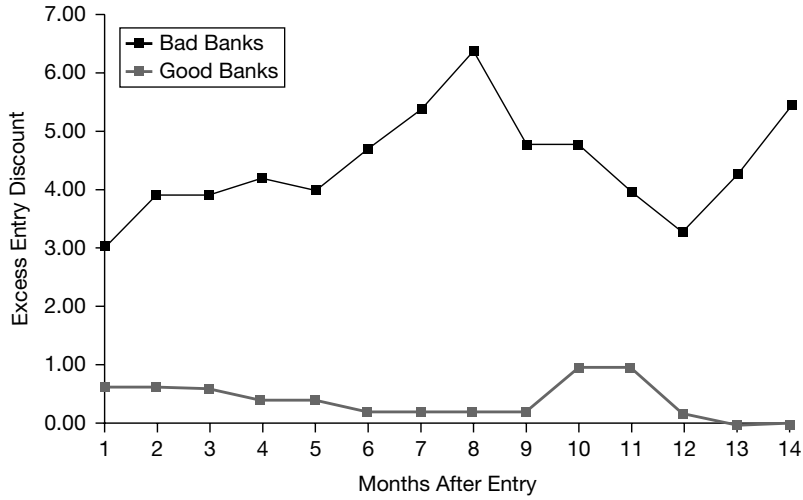


Figure 3.2 Excess entry discounts for good and bad banks: Tennessee

the notes of new banks relative to those of seasoned peers, participants could distinguish good banks from bad banks and (relatively) priced them accordingly.

As illustrations, figures 3.2 and 3.3 plot the average excess entry discounts (for the whole period) over the first year for the good banks and the bad banks for Tennessee and New York. It is clear that the good banks' excess entry discounts are lower initially and converge to zero by 1 year (by definition). The excess entry discounts of the bad banks diverge from the modal discount.

3.7.2. The Informational Basis of Distinctions between New Banks

What information could have led market participants to initially discriminate between entering new banks, more heavily discounting those that, in fact, did turn out to be insolvent? Part of the answer to this question is provided by table 3.8. Table 3.8 shows some average balance sheet ratios for banks in New York State. The data are divided between country banks and city banks since

an ordinary t-statistic, the following statistic was calculated:

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}.$$

This quantity does not follow the Student's t distribution when $\mu_1 = \mu_2$, but the degrees of freedom can be adjusted so that standard t tables can be used (see Snedecor and Cochran 1980). In both tables 3.7 and 3.8, the degrees of freedom shown are the adjusted degrees of freedom.

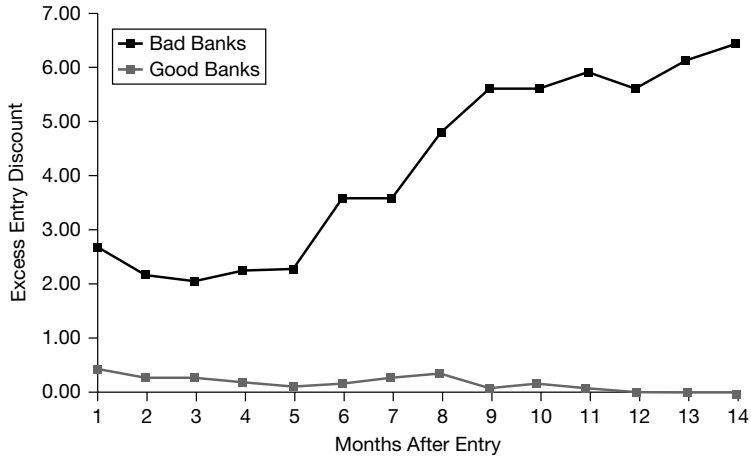


Figure 3.3 Excess entry discounts for good and bad banks: New York

these two groups have significantly different balance sheets. These data may have been available to market participants when the bank opened, and certainly were available by the end of the year, when the state regulatory authorities collected and published the data. On the liability side of the balance sheet, the mean ratios of notes to total assets, deposits to total assets, specie to total assets, and capital to total assets are computed. On the asset side, the ratios of real estate loans (mortgages) to total assets, loans and discounts to total assets, and stock to total assets are computed. (Omitted are such categories as due from banks, due to banks, etc.) Note that there were no new good city banks during the period.

As seen in table 3.8, there are several notable differences between the various groups of banks. With respect to city banks, bad banks have significantly *more* notes and stock (i.e., bonds) whereas they have significantly *fewer* deposits, *less* specie, and *less* real estate.[25] Deposits and real estate require some time to acquire market share, whereas stock can be easily purchased as an asset. When seasoned country banks are compared to bad (new) country banks, bad banks have significantly *more* deposits and stock whereas they have significantly *less* specie, *less* real estate, and *fewer* loans. Good (new) country banks have significantly *more* notes, specie, stock, and capital than seasoned country banks and have *fewer* deposits, *less* real estate, and *fewer* loans. Finally, when bad (new) country banks are compared to good (new) country banks, bad banks have significantly *more* deposits and stock and have *fewer* notes, *less* specie, and *less* real estate.

Recall that the model predicts that new banks can expect more notes to be redeemed since they are perceived as being riskier than seasoned banks; these

25. The term "stocks" refers to what we call bonds in modern parlance.

Table 3-8. COMPARISON OF BALANCE SHEET RATIOS FOR NEW YORK STATE BANKS

	New York City Banks						New York Country Banks					
	Seasoned Banks ($N = 687$)	New Bad Banks ($N = 63$)	t'	Seasoned Banks ($N = 3{,}003$)	New Bad Banks ($N = 104$)	t'	Seasoned Banks ($N = 3.003$)	New Good Banks ($N = 249$)	t' ($N = 104$)	New Bad Banks ($N = 249$)	New Good Banks ($N = 249$)	t'
Notes/total assets	.093 (.061)	.113 (.064)	–2.38*	.280 (.131)	.278 (.094)	.21	.280 (.131)	305 (.125)	–3.02*	.278 (.094)	.305 (.125)	–2.22*
Deposits/ total assets	.376 (.106)	.326 (.101)	3.74*	.180 (.111)	.209 (.110)	−2.64*	.180 (.111)	.149 (.091)	5.07*	.209 (.110)	.149 (.091)	4.91*
Specie/total assets	.092 (.053)	.070 (.040)	4.05*	.018 (.017)	.016 (.008)	2.37*	.018 (.017)	.022 (.029)	–2.15*	.016 (.008)	.022 (.029)	–3.00*
Real estate/total assets	.034 (.042)	.026 (.025)	2.26*	.024 (.032)	.012 (.017)	6.79*	.024 (.032)	.017 (.027)	3.87*	.012 (.017)	.017 (.027)	–2.09*
Loans/total assets	.616 (.102)	.591 (.150)	1.30	.569 (.173)	.517 (.140)	3.69*	.569 (.173)	.538 (.183)	2.58*	.517 (.140)	.538 (.183)	–1.17
Stock/total assets	.074 (.077)	.104 (.092)	–2.51*	.152 (.152)	.249 (.100)	−9.52*	.152 (.152)	.205 (.167)	–4.84*	.249 (.100)	.205 (.167)	3.05*
Capital/total assets	.365 (.105)	.465 (.097)	–7.78	.398 (.125)	.429 (.119)	–2.61	.398 (.125)	.439 (.112)	–5.50*	.429 (.119)	.439 (.112)	–.73

NOTE: Standard errors are in parentheses.
* Significant at the 5 percent level.

redemptions must be honored to avoid bankruptcy. What is clear from the comparisons above is that bad banks, whether city or country, have less specie reserves than seasoned banks or good banks. Since new bad banks' notes face significantly higher discounts, more of their notes would be redeemed than notes of new good banks. But their specie to total assets ratio is significantly lower than that of seasoned banks or new good banks. It appears that they are less able to honor redemptions. This is consistent with the redemption option allowing market participants to monitor banks and discover bank type quickly.

Table 3.8 examines each ratio individually. I next ask which balance sheet characteristics are priced by the market for new banks' notes. Table 3.9 addresses this by regressing the excess entry discounts for new banks in New York State on the balance sheet ratios. Because balance sheet ratios are often highly correlated, several specifications are examined. The only ratios that are significant are the ratios of notes to total assets and specie to total assets. As expected, market participants demanded higher excess entry discounts for banks with low amounts of specie (to total assets) and high amounts of notes (to total assets). It is perhaps surprising that the capital to total assets ratio is not important, but perhaps the reason is that it is a book value measure.[26]

3.8. TECHNOLOGICAL CHANGE AND PRIMARY NOTE PRICES

During the Free Banking Era, there was enormous technological change: the railroad and the telegraph were introduced and diffused across the United States. The railroad was introduced in England in the 1820s and spread to the United States shortly thereafter. Between 1838 and 1860, railroad mileage increased from 3,000 miles to over 30,000 miles (see Fogel 1964; Fishlow 1965). The first telegraph line was strung from Baltimore to Washington in 1846 and then from Philadelphia to New York. By 1860 there were 50,000 miles of telegraph lines. (The continent was spanned in 1861.) Five million messages per year were sent by telegraph in 1860 (see Thompson 1947; Du Boff 1980, 1983, 1984). These improvements affect the time it takes to return notes to an issuing bank and may have allowed more accurate predictions of a bank's type. In this section, I examine whether these technological changes affected the market for new banks' notes. In order to examine the effects of these technological changes, an index

26. There is also a timing problem. The date of the bank's entry according to *Van Court* is typically earlier than the regulatory authorities' publication of the balance sheet data. During this interval the market value of bank equity could change by a lot because of learning by market participants via redemptions.

Table 3-9. Determinants of Excess Entry Discounts: New York State ($N = 541$) Dependent Variable: Excess Entry Discount

Independent Variable	Parameter Estimate (1)	*t*-Value	Parameter Estimate (2)	*t*-Value	Parameter Estimate (3)	*t*-Value
Intercept	.0008	.225	.0004	.145	−.0031	−1.977
Deposits/total assets	−.005	−1.499	−.006	−1.780	...	...
Real estate/total assets	.006	.066	.011	1.230	.0134	−1.447
Loans/total assets	−.0006	−.220	...	...	...	...
Stock/total assets	−.005	1.410	...	...	...	...
Notes/total assets	.008	2.120	.005	1.600	.0082	−3.273
Specie/total assets	−.08	−7.870	−.080	−7.800	−.0803	−8.177
Capital/total assets	.002	.560	.002	.540	.0049	−1.978
R^2	.1855		.1814		.1766	
F-value	17.38		23.76		28.80	
Prob > *F*	.0001		.0001		.0001	

of technological change is required. Section 3.8.1 discusses the construction of such an index.

3.8.1. Measuring Technological Change

Indices of the time it took to get from Philadelphia to the largest city in each state or territory in the sample were constructed from pre–Civil War travelers' guides, which provided the most commonly used routes and the means of transport (steamship, canal boat, stagecoach, or railroad) along each leg of the trip. The guides also provide the number of miles traveled on each particular leg. This information was combined with estimates of the rate of travel (miles per hour) for each mode of transport to construct the index[27] (see Gorton [1989*a*] for details). The index was constructed for three years: 1836, 1849, and 1862 (the only years for which the travel guides could be located). These years correspond roughly to three regimes: 1839–45, 1846–50, and 1850–58. Prior to 1845, neither the railroad nor the telegraph had made much progress. Progress was made in the middle period and by the last period had become widespread.

The index does not explicitly account for the diffusion of the telegraph. However, since the telegraph tended to be strung alongside railroad tracks and the main innovation reducing travel time was the railroad, the index roughly captures the influence of both the railroad and the telegraph (see Thompson 1947).

Improvements in travel times were dramatic during the two decades from 1839 to 1858. Figure 3.4 shows the travel times for representative locations for each of the three years. It is important to note that there is a good deal of cross-section variation: for some locations the largest gains came in the middle period, whereas for others they came in the last period.

3.8.2. Reputation Formation and Technological Change

The introduction of the telegraph and the railroad should affect the pricing of new bank notes initially. There are two effects. First, monitoring via note redemptions takes time. Since technological change reduces the amount of time it takes to redeem a note, monitoring via redemptions will improve ceteris paribus. Second, initial estimates of new banks' types may improve.

27. Gorton (1989*a*) also computes the cost of a trip to each particular location. This is highly correlated with the time it takes, so here only the time to return to the issuing bank is analyzed.

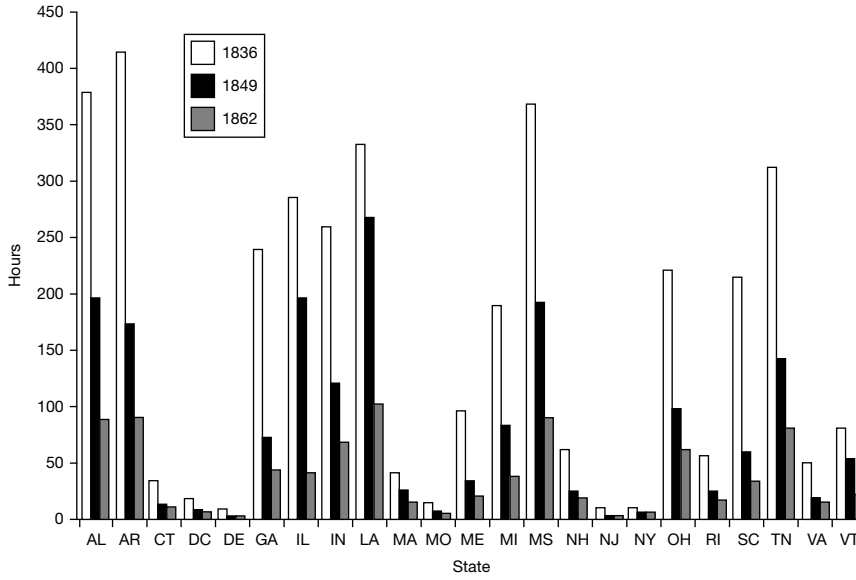


Figure 3.4 Technological change

As travel time falls, notes can be returned for redemption more quickly, allowing new banks to be monitored faster. While this would affect the prices of all banks' notes and hence the modal discounts, it would have a greater effect on initial note prices. A reduction in redemption time corresponds to a decrease in maturity. But, as argued above, this would reduce the prices of bad banks' notes by more than those of good banks' notes (see the lemma in Appendix A). Thus there is a greater incentive to redeem the notes of bad banks, and they would become insolvent faster ceteris paribus. If bad banks are detected faster, then the excess entry discount on the remaining banks' notes should be reduced faster (approaching the modal discount).

The second effect concerns the possibility of improved information about bank type initially. The telegraph, in particular, would allow information about a new bank's ability to redeem notes to have reached distant locations before the new bank's notes had arrived there. Organizing a new bank took time because either a charter had to be granted by the state legislature or a free bank had to establish itself with the regulatory authorities by depositing state bonds. There was, thus, an interval between the time in which a bank was established and the time of its first note issuance. During this period, information could flow to other parts of the country. With technological change, *Van Court*'s initial note prices may have become more accurate.[28]

28. The effects of improved estimates of σ on the note price are unclear: the option pricing formula is nonlinear in the variance so that an unbiased estimate of the variance does not produce

More accurate initial note prices should force the average quality of entering banks to improve. Average quality can improve if entering banks reduce their asset risk, reduce leverage, or hold more reserves, for the same asset risk. Recall that in Diamond's model there are three types of borrowers (i.e., banks): good banks, bad banks, and banks that can choose between good and bad projects. As time goes by, some borrowers default. On average, these will be bad banks. But a consequence of such defaults is that the rate charged to the survivors goes down, which can, in turn, cause the borrowers with a choice of projects to choose the safe project, further improving the average quality of the survivors. To the extent that market participants can detect bad types initially (and price them accordingly), the lower interest rate can be charged to the remaining banks on issue. But then the effect on those borrowers that can choose between projects is felt immediately, reducing the interest rate for the surviving banks. Thus the prediction of Diamond's model would be that improved information should cause the excess entry discount to decline with technological change. The argument also implies that market participants should be better able to distinguish between good and bad banks with technological change. We shall now examine these predictions.

3.8.3. Tests for Effects of Technological Change

The first prediction, that technological change should reduce the average excess entry discount, is examined in panel B of table 3.4 above. This panel computes the excess entry discount by subperiod. As can be seen in table 3.4, there is a marked decline in the mean excess entry discount, though it is still significantly positive in the last period. Table 3.7 addresses the second prediction, that technological change should improve initial information sets to allow market participants to distinguish good banks from bad banks. In table 3.7 the excess entry discount for good banks is insignificantly different from zero after 1845; market participants detect good banks at entry in the later two periods. These results suggest that the three periods are different but do not make use of the cross-section variation in improvements in technology captured by the travel time index.

Table 3.6 uses the travel time index and provides further evidence of the importance of technological change. Column 1 of the table includes the variable travel time, which is the index of the time of a trip back to the issuing bank, discussed above. The index is measured in hours. In the regression the three years for which the index is constructed were assigned to the three regimes. If travel

an unbiased estimate of the note price. The sign of the bias cannot be unambiguously determined (see Boyle and Ananthanarayanan 1977).

time falls, then, as explained above, the excess entry discount should fall. Indeed, the positive correlation is detected in the regression, again confirming the first prediction. To get some sense of the importance of the reduction in travel time, consider the mean excess entry discount for the entire period, .0697. If this corresponds to an average travel time of 3 days (72 hours) and this time is reduced to 1 day, then the excess entry discount falls to .0216, a third of the initial excess entry discount. Thus technological change is not only significant in the regression but quite dramatic in practical terms.

The rest of table 3.6 addresses the issue of whether the information possessed by market participants about new banks became finer over time. Columns 2–4 of table 3.6 present a time-series, cross-section seemingly unrelated regression of the excess entry discounts on the a priori risk factors for the three subperiods. Notably, the risk factors of state banking systems are priced in the early period, but in the last period they are not priced. In the early period, market participants know the characteristics of state banking systems and possibly little else about entering banks. But in the last period, excess entry discounts have fallen, though they are still significantly positive, and the market still distinguishes between good and bad banks, but the state characteristics are not priced. This would occur if market participants had finer information than state risk characteristics.

Technological change allowed market participants to have finer information about entering banks, imposing tougher discipline on entrants. Excess entry discounts declined as the time it took to transmit messages fell because of technological change. In fact, good banks' entry discounts were not insignificantly different from those of seasoned banks in the middle and late periods.[29]

3.9. CONCLUSION

Diamond's (1989) theory of reputation formation appears to accurately describe bank note issuance during the American Free Banking Era. The notes of new banks were more heavily discounted than the notes of banks with credit histories. Consumers, who use the bank notes as a medium of exchange, had an incentive to return the notes of higher-risk banks for redemption. This mechanism allowed consumers to learn quickly whether new banks had the appropriate asset risk. Redemption and reputation, combined with public and private restrictions on risk taking that limited the degree of adverse selection, explain the success of the Free Banking Era (in the sense that wildcat banking was not widespread).

29. In the analysis the degree of adverse selection was conceptually held constant. But the degree of adverse selection might be correlated with technological change. Though this correlation is not directly testable, it seems plausible that it would be causal; i.e., technological change reduced the degree of adverse selection.

APPENDIX A

This Appendix presents a simple model of bank notes, based on Svensson (1985) and Gorton (1993). The main simplification is that the model assumes that only privately issued notes can be used in exchange; the costs of using specie make notes preferable.

Assume that agents are identical except that they are spatially separated. Let d be a measure of the distance from an agent's home location to the distant market, where the agent trades at time t. (A time subscript on d will usually be omitted for ease of notation.) Because of symmetry, the distance measure, d, is an index of agents' locations. (The home location is $d = 0$.) The representative agent (at a representative location) is assumed to prefer goods procured from locations further from home rather than nearer to home. The agent's objective is to maximize

$$E_t \left[\sum_{j=t}^{\infty} \beta^{j-t} U(C,d) \right], \qquad \textbf{(3A.1)}$$

where C is consumption, $0 < \beta < 1$, $U'_C > 0$, $U''_{CC} < 0$, $U'_d > 0$, and $U''_{dd} < 0$. The assumption that utility depends on distance says that the "same" good purchased further away "tastes" better; it is intended to capture the notion of a division of labor, motivating trade. Each agent is endowed with a non-tradable project that returns a random amount at date t, $y_t(d)$, of a single nonstorable consumption good. Endowments are independently, identically, lognormally distributed at each date and location. Assume that the current endowment, $y_t(d)$ (each location d), is public information.[30] Expectations below are taken over uncertainty concerning future endowments. The standard deviation of endowments at location d is $\sigma(d)$ and is assumed constant through time. Later, however, we shall briefly consider thought experiments in which an agent a distance d away has a higher $\sigma(d)$ than other agents at that location, and also the case in which σ may be chosen by the agent.

Since agents prefer goods from distant locations, they will trade. Assume that agents face a cash-in-advance constraint that can be satisfied only by issuing private money. Each agent issues two types of claims against future endowments: bank notes and equity. The notes are non-interest-bearing debt claims that allow for conversion into consumption goods on demand at par at the location of the issuing agent. For simplicity the equity does not pay dividends.

Each agent is to be thought of as a buyer-seller pair, as in Lucas (1980). There is a division of labor between the household seller and the household

30. Each location d receives the same endowment, suggesting the interpretation of the randomness as a geographical weather shock. Such information was widely reported in newspapers and by travelers.

buyer. Each household will be involved in transactions at two locations each period, corresponding to this division of labor. At the home location, the seller stays at home and sells the household endowment (minus the amount of notes that the household has redeemed, explained below) to buyers from other locations, receiving bank notes of other agents in exchange. The seller receives notes with a value equal to y_t (minus the amount of notes that the household has redeemed). Also, at the home location, the household trades in the securities market later. Notes at the home location are indexed by *d,* indicating the distance to the issuing bank from the home location. Indicate note prices (in terms of consumption units) at the home location of notes issued by banks a distance d away by $P_t(d)$. The other transaction is carried out by the buyer and occurs at a distant location. Only one (distant) market can be visited at each date *t.* The buyer chooses a distance, *d,* and a direction to travel, and purchases goods at that distant location, paying for them with bank notes.[31] We shall also need to index notes at this location. Let d' be the distance to the issuing bank from the distant location at which the buyer purchases goods. Indicate note prices (in terms of consumption units) at the distant location of notes issued by banks a distance d' away by $P_t(d')$. Note that d' depends on d (though this dependence is suppressed). For example, $d' < d$ when the buyer goes to a distant market, which brings the bank note closer to the issuing bank. When the buyer goes to a distant market, which takes the note even further from the issuing bank, $d' > d$.

The sequence of events in period t is as follows. First, households receive their endowments $y_t(d)$. Second, households honor notes turned in for redemption (this is described below). Third, the markets for goods open. The buyer travels to a distant market carrying the portfolio of bank notes held over from the previous period and purchases C_t consumption units from sellers at that location, using bank notes, and then returns home. Simultaneously, the household seller sells goods remaining from the household endowment (after notes have been honored) in the home market, receiving bank notes in exchange. Fourth, households go to the securities market at their home location to trade bank notes and

31. The direction and distance the buyer will travel can be taken as certain. By symmetry, the direction the buyer travels in does not matter, though it will be taken into account when the household chooses a portfolio of notes to be carried over to finance consumption. The household will buy the notes of that distant location (d') in its home market in order to carry them to their home location, where they will trade at par, or at least at a lower discount. In this securities market at the home location, the notes will be sold at discounts. An alternative assumption is that the direction the buyer goes in is random and only the distance is chosen. In this case, the buyer will be forced to carry notes to a distant location, and they will be sold at discounts. The assumption of a random direction requires that this uncertainty be taken into account. The first assumption avoids this complication without changing the conclusions.

bank shares. Households choose a portfolio of notes and shares and, in particular, may decide to redeem some notes. The choice of the new portfolio of notes will reflect the direction and distance that the buyer will travel next period (this is currently known). Finally, consumption occurs and period t ends.

In order to give meaning to the notion of distance, assume that a note issued by an agent a distance d away takes d periods to return for redemption. Thus there is assumed to be an asymmetry between buyers and sellers. Buyers can carry a note a distance d in a single period, but a seller who receives the note requires d periods to receive the (risky) payoff to redeeming the note (if redemption of the note is chosen). This asymmetry is introduced for tractability.

Recall that $P_t(d')$ is the price (in terms of consumption units) of bank notes carried by the representative agent and traded at a location a distance d away at time t (d' is the distance from the market the buyer has chosen to the issuing bank). The cash-in-advance constraint faced by the buyer is

$$C_t \leq \sum_d P_t(d')N_{t-1}(d). \qquad \textbf{(3A.2)}$$

Each period the household may choose to send some notes for redemption at distant banks. The household may also face a demand for redemptions of its own notes. Redemptions are honored out of the household endowment before the markets for goods open. Let $N_t^R(d)$ be the amount of notes of banks at location d that are sent for redemption in period t. Notes that the household sent for redemption k periods ago will be honored this period if $d = k$. Otherwise, $d > k$, and the notes are still in transit.[32] The face amount the household must itself currently honor is $N_t^R(0)$.

When notes are redeemed, they are redeemed at face value if the bank is solvent. Otherwise, there is a loss. Let $P_t^R(d)$ be the price at which a note is redeemed; $P_t^R(d) = 1$ if the bank is solvent.[33] There are no bankruptcy costs, and the household is assumed to subsequently issue new notes with a face value equal to the face value of the amount redeemed.[34] For simplicity assume that no new equity is issued. Thus leverage is constant.[35]

32. Notes sent for redemption at time t will be *in transit* for d periods. Consequently, at any time t there may be notes sent for redemption in the past that have not been redeemed yet. This complication is dealt with by Gorton (1993) and, for simplicity, is ignored here.

33. The price $P_t^R(0) = \min[1, y_t/N_t^R(0)]$, where $N_t^R(0)$ is the face value of the notes that the household must honor this period.

34. A household cannot issue new notes in order to cover losses on old notes.

35. This can be viewed as a binding capital requirement.

Trading in the security market and the sending of notes for redemption occur at the home location. Let $q_t(d)$be the price of equity claims and $Q_t(d)$the number of shares of bank d stock held at time t. The household budget constraint is

$$\begin{aligned} P_t(0)^R N_t^R(0) + C_t + \sum P_t(d)N_t(d) + \sum P_t(d)N_t^R(d) + \sum q_t(d)Q_t(d) \\ \leq \sum P_t(d)N_{t-1}(d) + \sum q_t(d)Q_{t-1}(d) + y_t \\ + \sum_{d=k} P_t^R(d)N_{t-k+d}^R(d) + P_t(0)N_t^R(0). \end{aligned}$$

The right-hand side of the inequality lists the sources available to the household. They consist of, respectively, notes held over from the previous period, the equity portfolio held over, the household endowment, redemptions received, and new notes issued. These sources are used to finance the items on the left-hand side: the amount of the household's own notes that are redeemed, consumption, a new portfolio of notes, notes sent for redemption, and an equity portfolio. Rewriting the budget constraint, we get

$$\begin{aligned} C_t \leq \sum P_t(d)\left\{N_{t-1}(d) - \left[N_t(d) + N_t^R(d)\right]\right\} + \sum q_t(d)\left[Q_{t-1}(d) - Q(d)\right] \\ + y_t + \sum_{d=k} P_t^R(d)N_{t-1}^R(d) + N_t^R(0)\left[P_t(0) - P_t^R(0)\right]. \end{aligned} \quad \textbf{(3A.3)}$$

The representative agent chooses a distance to travel in period t, d; an amount of notes of each type, d, to be sent for redemption, $N_t^R(d)$; an amount of notes of each type, d, to be used to satisfy the cash-in-advance constraint, $N_t(d)$; and an amount of equity shares of each type, $Q_t(d)$, to maximize (3A.1), subject to (3A.2) and (3A.3). The first-order conditions for distance to travel (d), the amount of each note type to redeem $(N_t^R(d))$, the amount of each note type to hold$(N_t(d))$, and the amount of each equity type to hold $(Q_t(d))$ are, respectively,

$$U'_{dt} = -E_t\left\{\mu_t \sum_d \frac{\partial P_t(d')}{\partial d}[N_{t-1}(d)]\right\}. \quad \textbf{(3A.4)}$$

$$U'_{Ct} \geq \beta^d E_t\left[U'_{Ct+d}\frac{P_{t+d}(d)}{P_t(d)}\right] \quad \text{each } d, \quad \textbf{(3A.5)}$$

$$U'_{Ct} \geq \beta E_t\left\{U'_{Ct+1}\left[\frac{P_{t+1}(d)}{P_t(d)}\right] + \mu_{t+1}\left[\frac{P_{t+1}(d')}{P_t(d)}\right]\right\} \quad \text{each } d, \quad \textbf{(3A.6)}$$

and

$$U'_{Ct} = \beta E_t \frac{[U'_{Ct+1} qt + 1'(d)]}{q_t(d)} \quad \text{each } d, \tag{3A.7}$$

where E_t indicates the expectation conditional on information available at time t, and μ is the Lagrange multiplier associated with the cash-in-advance constraint. (There is also a transversality condition for the notes of each bank.)

Equilibrium requires that (1) the goods market clear at each location d, $C_t(d) = y_t(d) - P_t^R(d)N_t^R(d)$; (2) the equity market clear at each location d, $Q_t(d) = Q_{t-1}(d) = 1$; and (3) the note market clear at each location d, $N_t(d) + N_t^R(d) = N_{t-1}(d)$. Condition 1 determines prices of notes at each location. Conditions 2 and 3 determine security prices for bank equity and notes issued by distant banks.

In the securities market, an agent faces a choice between holding a particular bank note for another period to satisfy the cash-in-advance constraint (eq. [3A.6]) and sending the note back to the issuing agent for redemption, resulting in a risky payoff in d periods (eq. [3A.5]). If (3A.5) and (3A.6) are satisfied with equality, the agent must be indifferent between these alternatives. In particular, if (3A.5) holds as an equality, then the notes can be priced as risky pure discount bonds with maturity d.[36] Further, if preferences display constant relative risk aversion, then a closed-form solution for note prices based on Black and Scholes (1973) can be derived. (The proof of this proposition is standard and is due to Rubinstein [1976].)[37] The price of a note is then given by

$$P_t(d) = \left[N_t^R(d)\right]^{-1} \left\{ V_t(d) \left[1 - N\left(h_D + \sigma\right)\right] + (1 + r_f)^{-1} D_t^R(d) N(h_D) \right\},$$

where

$$h_D \equiv \frac{\ln\left[V_t(d)/N_t^R(d)\right] + \ln(1 + r_f)}{\sigma} - \frac{\sigma}{2},$$

σ is the standard deviation of one plus the rate of change of the value of the bank (i.e., the standard deviation of output), r_f is the risk-free rate of interest (assumed constant), $V_t(d)$ is the value of the debt and equity claims on household d at time t, and $N(\cdot)$, without a superscript, indicates the cumulative normal distribution function.[38]

36. If no notes are sent for redemption, then (3A.5) does not hold as an equality, but provides a bound on the note price. The remaining case occurs when the bank's notes are sent for redemption so that (3A.5) holds with equality but (3A.6) does not: i.e., the notes are more valuable being redeemed than they are being used as a means of exchange next period.

37. This assumes that there are no notes currently in transit.

38. For simplicity the model has no riskless security. However, the shadow price of a riskless bond can always be calculated. A riskless security could easily be incorporated.

This pricing formula is useful because it shows that the value of a note, $P_t(d)$, varies inversely with time to maturity (d), risk (σ) and leverage (see Merton 1974). Note, in particular, that the value of the note is decreasing in maturity, d.

Condition (3A.4) determines how far the buyer should choose to travel. By symmetry, the direction the buyer travels in is irrelevant (this was chosen before trading in the securities market and is currently known). Consider a buyer traveling to a distant location that takes a note even further away from the issuing bank than the home location (i.e., $d' > d$). In that case, maturity is increasing since it will take longer to return from the buyer's market. From the pricing formula we know that in this case $\partial P_t(d')/\partial d < 0$; that is, notes decline in value as they travel further away from the issuer. On the other hand, at the distant location the buyer is going to, some notes will be closer to the issuing bank, so maturity will have declined for these notes, and $\partial P_t(d')/\partial d < 0$. No matter what direction the buyer travels in, some notes in his portfolio will increase in value (as he moves closer to the issuing bank) and some notes will decline in value (as he moves further away from the issuing bank). According to (3A.4), the optimal distance to travel is chosen to equate the marginal benefit of increased distance (in terms of the goods' tasting better) to the marginal cost, which is the capital loss associated with carrying the notes further away from home and, hence, being able to purchase less.

The model above considers a setting in which all banks (households) at each location have access to the same project. In order to address the issue of new banks without repeating the work of Diamond (1989), consider allowing a new bank to enter the market at a given location. Assume that this new bank is perceived by other households to be of higher risk, $\sigma_N > \sigma_S$, where σ_S is the variance of the seasoned banks' project return (at location d). The new bank is the same as the seasoned banks at its location except with respect to project risk. I shall show that in equilibrium the notes of the new bank (N) will be redeemed, enforcing the equilibrium in which all banks have the risk of the seasoned banks (S) (taken as exogenous).

The following lemma is a standard result from contingent claims (see Merton 1974).

LEMMA. Consider two banks, bank N (for new) and bank S (for seasoned), which are the same distance away (d) and have the same leverage, but have different risk. In particular, $\sigma_N > \sigma_S$, so $P_t^S(d) > P_t^N(d)$. Then

$$\frac{\partial P_t^N(d)}{\partial d} < \frac{\partial P_t^S(d)}{\partial d}.$$

The lemma says that the value of bank N's notes decays at a faster rate as the distance away from the bank is increased. Note that the optimal choice of distance using the new bank's notes, d_N, is lower than the optimal choice of distance using the seasoned banks' notes, $d_S (d_N < d_S)$, because $\sigma_N > \sigma_S$. We can now state the following proposition.

PROPOSITION A1. If the notes of two banks at the same distant location (d), with identical amounts of notes outstanding and identical leverage, circulate to the same extent at a particular location, then they must have identical risk; that is, the two banks have the same σ's.

Proof. The proposition is proved by contradiction. Consider two banks, bank S and bank N, identical except that $\sigma_N > \sigma_S$. I shall show that the notes of bank N will tend to be sent for redemption, whereas those of bank S will circulate (i.e., be used to satisfy the cash-in-advance constraint). Let N_t^{Ri} be the amount of bank i's notes being sent for redemption and let N_t^i be the amount of bank i's notes being held for circulation, $i = N$ or S. Suppose that both types of notes circulate to the same extent and that the household sends the same amount of each for redemption. I shall show that this cannot be an equilibrium. If both types of notes circulate, then $N_t^i > 0$ for $i = S, N$ and (3A.6) holds with equality for each bank's notes. Also, by hypothesis (of an interior solution), (3A.5) holds with equality for each note type, that is, $N_t^{Ri} > 0$ for $i = S, N$.

To show that this cannot be an equilibrium, consider the following rearrangement of the agent's portfolio. Reduce the amount of bank S notes being sent for redemption by ΔN_t^{RS}, increasing the amount of bank S notes being held for circulation by the same amount. Increase the amount of bank N notes being sent for redemption by $(P_t^N / P_t^S) \Delta N_t^{RN} = \Delta N_t^{RS}$, so that the expected value of the total amount being sent for redemption is the same. (Note that this strategy is self-financing since $P_t^N \Delta N_t^{RN} = P_t^S \Delta N_t^{RS}$.) Then, with respect to the expected value of future redemptions, the agent is no worse off. But the amount of bank S notes being held for circulation is greater and the amount of bank N notes being held for circulation is decreased. Now, using (3A.8), consider the effect on the choice of distance:

$$\Delta U_d = -E\left[\mu_t \left(\frac{\partial P_t^S}{\partial d} \Delta N_t^{RS} - \frac{\partial P_t^N}{\partial d} \Delta N_t^{RN} \frac{P_t^S}{P_t^N}\right)\right].$$

But, imposing that the strategy is self-financing, recalling that $P_t^S > P_t^N$, and noting that the difference in partial derivatives is negative (by the lemma), we see that the agent is better off. Q.E.D.

Finally, consider the case of endogenous asset risk, that is, an "out-of-equilibrium" wildcat bank that increases asset risk above σ_S. Suppose that a new bank issues notes for the first time at date t. These notes, printed at date $t-1$, will be used to finance initial consumption so that $C_t \leq P_t(d)N_{t-1}$ is the initial budget constraint and, coincidentally, the cash-in-advance constraint; N_{t-1} is the initial amount of notes printed. Next period this agent/bank will have none of its own notes (since they will have been spent at a distant location) but will have received other agents' notes and will have its own bank equity, which can be used to finance consumption. The first-order condition for choice of risk, σ, is

$$-U'_{Ct}\frac{\partial P_t}{\partial \sigma}N_{t-1} = \beta E_t\left[U'_{Ct+1}\frac{\partial q_{t+1}}{\partial \sigma}(Q_t - Q_{t+1})\right].$$

Since $\partial q_{t+1}/\partial \sigma > 0$, the increase in risk results in a higher value of the bank equity (i.e., equity is valued as a call option on the value of the bank in the standard way). Selling this equity next period will allow the wildcat bank to realize the benefits of increased risk.[39] But the cost of the increase in risk is that $\partial P_t/\partial \sigma < 0$; that is, a smaller amount of consumption can be purchased when the notes are carried to a distant market initially to get them into circulation. In other words, market participants, recognizing the incentives of the bank, will discount its notes appropriately, penalizing the bank when it first introduces the notes into the market. Consequently, this bank will not choose an infinite amount of risk.

A wildcat bank chooses a level of risk higher than σ_S. In that case, if the arbitrage bound is violated, all its notes will be redeemed, say, next period.[40] Then the wildcat bank can benefit only if it does not go bankrupt and the choice of risk is given by

$$-U'_{Ct}\frac{\partial P_t}{\partial \sigma}N_{t-1} = \beta\int_0^{y^*}\left[U'_{Ct+1}\frac{\partial q_{t+1}}{\partial \sigma}(Q_t + Q_{t+1})\right]f(y)dy,$$

where $y^* = N^R$, indicating the level of output at which the bank is bankrupt when $N^R(=N_{t-1})$ notes are redeemed. Thus the equilibrium in which all banks choose σ_S is supported if adding more risk cannot satisfy the first-order condition above. In that case, the threat of redemption prevents wildcat banking.

39. Of course, in equilibrium the representative household must hold all the equity and could not benefit by selling it.

40. In other words, since other market participants understand the incentives of the wildcat bank, $d = 1$, which means that all the wildcat bank's notes will be redeemed next period.

APPENDIX B

Table 3-B1. COVERAGE OF VAN COURT'S BANK NOTE REPORTER: STATES AND DATES

States with Complete Coverage, February 1839–December 1858		**States with Incomplete Coverage***		**States Listed as "Uncertain" or Not Listed**
United States	**Canada**	**United States**	**Canada**	
Alabama	Canada†	Arkansas (1840–58)	New Brunswick (1840–48)	Iowa Territory
Connecticut	Nova Scotia	Florida (1842–58)		Minnesota
Delaware		Illinois (July 1856–58)		Missouri
District of Columbia		Indiana (1857)		Texas
Georgia		Michigan (1853)		
Kentucky		Mississippi (1839, 1841–43, 1852–58)		
Louisiana		Nebraska (1840–47)		
Maine		New Hampshire (1857–58)		
Maryland		Virginia (1846–47, 1853–54)		
Massachusetts		Wisconsin (1839–55)		
Montana‡				
Pennsylvania				
New Jersey				
New York				
North Carolina				
Ohio				
Rhode Island				
South Carolina				
Tennessee				
Vermont				

* Incomplete coverage means that the *Van Court's Bank Note Reporter* did not quote a price for banks in that state for that month. The state may have been listed, though, and the notes of the banks in that state described as "all uncertain." Dates in parentheses indicate periods for which the data were missing.

† Canada includes banks located in provinces other than Nova Scotia or New Brunswick.

‡ Montana became the forty-first state in 1889.

REFERENCES

Asquith, Paul, Mullins, David W., Jr., and Wolff, Eric D. "Original Issue High Yield Bonds: Aging Analyses of Defaults, Exchanges, and Calls." *J. Finance* 44 (September 1989): 923–52.

Atherton, Lewis E. *The Frontier Merchant in Mid-America.* Columbia: Univ. Missouri Press, 1971.

Black, Fischer, and Scholes, Myron S. "The Pricing of Options and Corporate Liabilities." *J.P.E.* 81 (May/June 1973): 637–54.

Boyle, Phelim P., and Ananthanarayanan, A. L. "The Impact of Variance Estimation in Option Valuation Models." *J. Financial Econ.* 5 (December 1977): 375–87.

Calomiris, Charles W. "Deposit Insurance: Lessons from the Record." *Econ. Perspectives* [Fed. Reserve Bank Chicago] 8 (May/June 1989): 10–30.

Calomiris, Charles W., and Schweikart, Larry. "Was the South Backward? North-South Differences in Antebellum Banking during Normalcy and Crisis." Manuscript. Urbana: Univ. Illinois, Dept. Finance, 1988.

Carothers, Neil. *Fractional Money: A History of the Small Coins and Fractional Paper Currency of the United States.* New York: Wiley, 1930. Reprint. New York: Kelley, 1967.

Cleaveland, John. *The Banking System of the State of New York.* New York: Voorhies, 1857.

Dewey, Davis R. *State Banking before the Civil War.* Washington: Government Printing Office, 1910.

Diamond, Douglas W. "Reputation Acquisition in Debt Markets." *J.P.E.* 97 (August 1989): 828–62.

____. "Monitoring and Reputation: The Choice between Bank Loans and

Directly Placed Debt." *J.P.E.* 99 (August 1991): 689–721.

Dillistin, William H. *Bank Note Reporters and Counterfeit Detectors, 1826–1866.* Numismatic Notes and Monographs, no. 114. New York: American Numismatic Soc., 1949.

Du Boff, Richard B. "Business Demand and the Development of the Telegraph in the United States, 1844–1860." *Bus. Hist. Rev.* 54 (Winter 1980): 459–79.

____. "The Telegraph and the Structure of Markets in the United States, 1845–1890." In *Research in Economic History,* vol. 8, edited by Paul Uselding. Greenwich, Conn.: JAI, 1983.

____. "The Telegraph in Nineteenth-Century America: Technology and Monopoly." *Comparative Studies Society and Hist.* 26 (October 1984): 571–86.

Ederington, Louis H. "The Yield Spread of New Issues of Corporate Bonds." *J. Finance* 29 (December 1974): 1531–43.

Fama, Eugene F. "Financial Intermediation and Price Level Control." *J. Monetary Econ.* 12 (July 1983): 7–28.

Fishlow, Albert. "Antebellum Interregional Trade Reconsidered." *A.E.R. Papers and Proc.* 54 (May 1964): 352–64.

____. *American Railroads and the Transformation of the Ante-Bellum Economy.* Cambridge, Mass.: Harvard Univ. Press, 1965.

Fogel, Robert W. *Railroads and American Economic Growth: Essays in Econometric History.* Baltimore: Johns Hopkins Press, 1964.

Fung, W. K. H., and Rudd, Andrew. "Pricing New Corporate Bond Issues: An Analysis of Issue Cost and Seasoning Effects." *J. Finance* 41 (July 1986): 633–43.

Glasner, Lynn. *Counterfeiting in America.* New York: Potter, 1960.

Gorton, Gary B. "Ante Bellum Transportation Indices." Manuscript. Philadelphia: Univ. Pennsylvania, Wharton School, 1989. (*a*)

Gorton, Gary B. "An Introduction to Van Court's Bank Note Reporter and Counterfeit." Manuscript. Philadelphia: Univ. Pennsylvania, Wharton School, 1989. (*b*)

____. "Pricing Free Bank Notes." Manuscript. Philadelphia: Univ. Pennsylvania, Wharton School, 1993.

Gouge, William M. *A Short History of Paper Money and Banking in the United States.* Philadelphia: Ustick, 1833.

Grant, James. *A Treatise on the Law Relating to Bankers and Banking.* Philadelphia: Johnson, 1857.

Green, George D. *Finance and Economic Development in the Old South: Louisiana Banking, 1804–1961.* Stanford, Calif.: Stanford Univ. Press, 1972.

Hammond, Bray. *Banks and Politics in America, from the Revolution to the Civil War.* Princeton, N.J.: Princeton Univ. Press, 1957.

Kahn, James A. "Another Look at Free Banking in the United States." *A.E.R.* 75 (September 1985): 881–85.

Klein, Benjamin. "The Competitive Supply of Money." *J. Money, Credit and Banking* 6 (November 1974): 423–53.

Lindstrom, Diane L. "Demand, Markets, and Eastern Economic Development: Philadelphia, 1815–1840." *J. Econ. Hist.* 35 (March 1975): 271–73.

Lindvall, John R. "New Issue Corporate Bonds, Seasoned Market Efficiency and Yield Spreads." *J. Finance* 32 (September 1977): 1057–67.

Lucas, Robert E., Jr. "Equilibrium in a Pure Currency Economy." In *Models of Monetary Economies,* edited by John H. Kareken and Neil Wallace. Minneapolis: Fed. Reserve Bank, 1980.

Mercer, Lloyd J. "The Antebellum Interregional Trade Hypothesis: A Reexamination of Theory and Evidence." In *Explorations in the New Economic History: Essays in Honor of Douglass C. North,* edited by Roger Ransom, Richard Sutch, and Gary M. Walton. New York: Academic Press, 1982.

Merritt, Fred D. *The Early History of Banking in Iowa.* Iowa City: Univ. Iowa Press. 1900.

Merton, Robert C. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *J. Finance* 29 (May 1974): 449–70.

Mullineaux, Donald J. "Competitive Monies and the Suffolk Bank System: A Contractual Perspective." *Southern Econ. J.* 53 (April 1987): 884–98.

Pred, Allan R. *Urban Growth and City Systems in the United States, 1840–1860.* Cambridge, Mass.: Harvard Univ. Press, 1980.

Rockoff, Hugh. "Money, Prices, and Banks in the Jacksonian Era." In *The Reinterpretation of American Economic History,* edited by Robert W. Fogel and Stanley L. Engerman. New York: Harper and Row, 1971.

____. "The Free Banking Era: A Reexamination." *J. Money, Credit and Banking* 6 (May 1974): 141–67.

____. *The Free Banking Era: A Re-examination.* New York: Arno, 1975.

____. "New Evidence on Free Banking in the United States." *A.E.R.* 75 (September 1985): 886–89.

____. "Lessons from the American Experience with Free Banking." Working Papers on Historical Factors in Long-Run Growth, no. 9. Cambridge. Mass.: NBER, December 1989.

Rolnick, Arthur J., and Weber, Warren E. "Free Banking, Wildcat Banking, and Shinplasters." *Fed. Reserve Bank Minneapolis Q. Rev.* 6 (Fall 1982): 10–19.

____. "New Evidence on the Free Banking Era." *A.E.R.* 73 (December 1983): 1080–91.

____. "The Causes of Free Bank Failures: A Detailed Examination." *J. Monetary Econ.* 14 (November 1984): 267–91.

____. "Explaining the Demand for Free Bank Notes." *J. Monetary Econ.* 21 (January 1988): 47–71.

Rubinstein, Mark. "The Valuation of Uncertain Income Streams and the Pricing of Options." *Bell J. Econ.* 7 (Autumn 1976): 407–25.

Schweikart, Larry. *Banking in the American South from the Age of Jackson to Reconstruction.* Baton Rouge: Louisiana State Univ. Press, 1987.

Shepard, A. K. "A National Currency." *Merchants' Mag. and Commercial Rev.* 50 (January 1864): 15–17.

Smith, Walter B., and Cole, Arthur H. *Fluctuations in American Business, 1790–1860.* Cambridge, Mass.: Harvard Univ. Press, 1935.

Snedecor, George W., and Cochran, William G. *Statistical Methods.* 7th ed. Ames: Iowa State Univ. Press, 1980.

Sorensen, Eric H. "On the Seasoning Process of New Bonds: Some Are More Seasoned than Others." *J. Financial and Quantitative Analysis* 17 (June 1982): 195–208.

Svensson, Lars E. O. "Money and Asset Prices in a Cash-in-Advance Economy." *J.P.E.* 93 (October 1985): 919–44.

Thompson, Robert L. *Wiring a Continent: The History of the Telegraph Industry in the United States, 1832–1866.* Princeton, N.J.: Princeton Univ. Press, 1947.

Wasserfallen, Walter, and Wydler, Daniel. "Underpricing of Newly Issued Bonds: Evidence from the Swiss Capital Market." *J. Finance* 43 (December 1988): 1177–91.

Weinstein, Mark I. "The Seasoning Process of New Corporate Bond Issues." *J. Finance* 33 (December 1978): 1343–54.

Whitney, David R. *The Suffolk Bank.* Cambridge, Mass.: Riverside, 1878.

4

Pricing Free Bank Notes

GARY B. GORTON*

4.1. INTRODUCTION

In this paper I study the pricing of private money issued by banks prior to the Civil War. These bank notes were perpetual, risky, non-interest-bearing, debt claims with the right to redeem on demand at par in specie. Between 1838 and 1863, the Free Banking Era, thousands of different bank notes circulated, constituting the overwhelming bulk of the securities market during this period. Taking account of the redemption option, I show how the value of these private money contracts depends on state-specific risk factors and on the technological ability of market participants to travel back to issuing banks to redeem notes. The pricing model is then tested to determine whether note prices reflect these factors.

Private money contracts have traditionally been viewed as very difficult to enforce. The basic critique of private money issuance has been articulated by Milton Friedman (1959):

> . . . the contracts in question are peculiarly difficult to enforce and fraud peculiarly difficult to prevent . . . individuals may be led to enter into contracts with persons far removed in space and acquaintance, and a long period may elapse between the issue of a promise and the demand for its

* The comments and suggestions of the Penn Macro Lunch Group, participants at the NBER Meeting on Credit Market Imperfections and Economic Activity, the NBER Meeting on Macroeconomic History, and participants at seminars at Ohio State, Yale, London School of Economics and London Business School were greatly appreciated. The research assistance of Sung-ho Ahn, Chip Bayers, Eileen Brenan, Lalit Das, Molly Dooher, Henry Kahwaty, Arvind Krishnamurthy, Charles Chao Lim, Robin Pal, Gary Stein, and Peter Winkelman was greatly appreciated. This research was supported by National Science Foundation grant no. SES86-18130 and a University of Pennsylvania Research Fund grant for which the Author is very grateful. Versions of this paper previously circulated under other titles.

> fulfillment . . . A fiduciary currency ostensibly convertible into the monetary commodity is therefore likely to be overissued from time to time and convertibility impossible. Historically, this is what happened under so-called 'free banking' in the United States and under similar circumstances in other countries. (p. 6)

Friedman is referring to "wildcat banks," banks that opened and then inflated their currency to the point where it could not be continuously redeemed.[1] The banker then absconded with the proceeds, leaving the private money worth less than par. The result was, possibly large, losses to the note holders. Indeed, examining the American Free Banking period, Cagan (1963) estimated that note holders suffered losses on their note holdings of 25% per year. According to Rockoff (1975) losses on notes ranged from 7 cents on the dollar in Indiana to 63 cents per dollar in Minnesota.[2] On the other hand, Rockoff (1971, 1974a, 1974b, 1975, 1989) argues that wildcat banking appears to characterize the experiences of only some states. Rolnick and Weber (1982, 1983, 1984) examined the timing of bank closings in four free bank states (Minnesota, Indiana, Wisconsin and New York), arguing that free bank failures and losses were not due to systematic wildcat banking, but to recessions.

The consensus seems to be that wildcat banking was not a prevalent feature of the ante bellum banking system. This view is based on an examination of *ex post* evidence concerning the incidence of bank failures and losses across different state banking systems. Using price data, I focus on the question of whether market participants priced the risk of bank notes *ex ante.* The idea is the familiar one that market participants may well have understood the risks inherent in private money and priced them correctly. This is important for addressing the question of *why* there was so little wildcat banking.

Wildcat banking may have been prevented because private institutional arrangements and state regulations constrained banks effectively. Another important consideration is the design of the bank note contract. Given the constraints of available data I concentrate on these two issues in analyzing the pricing of bank notes. First, I ask whether bank note prices reflect private institutional and state regulatory factors that independent evidence suggests were important determinants of risk. State banking systems varied in allowing branch banking, in providing state insurance, and in allowing "free banking" in that entry into banking was less restrictive. (Free banking states required the deposit of state bonds against money issuance. Chartered banking states required a license from the legislature to operate, and imposed reserve and capital requirements.)

1. Friedman has apparently changed his views. See Friedman (1986, 1987) and Friedman and Schwartz (1986).

2. Knox (1903, p. 315) estimates the losses to note holders to have been "about 5% per annum."

Also banks in some states were members of formal or informal private bank associations which regulated members.

Secondly, to analyze bank note prices I take account of the redemption option in bank note contracts. This option may have been important in limiting bank risk-taking because it allowed note holders to run on banks which began to increase their risk, for example, by printing money. Pricing this embedded option requires taking account of the fact that to exercise the option a noteholder must return the note to the issuing bank. Returning the note to the issuing bank required using the available transportation technology. Indeed, Friedman's critique appears to be rooted in such considerations of technology; that is, if the bank is too far away then risk-taking cannot be effectively prevented. Others have also argued that the US was so technologically underdeveloped in this period that it was difficult to price the notes. Taylor (1951, p. 312) writes: "As long as transportation and communication were relatively slow and no effective clearing system had developed, mere distance from the centers of commerce was a valuable asset to a bank."

In pre-Civil War America communication and transportation were difficult, but dramatic change did occur. The introduction of the railroad drastically lowered transportation costs as it spread across the country during this period. Introduced in England in the 1820s, the railroad was quickly adopted in the US. Between 1838 and 1860 railroad mileage nationwide increased from about 3,000 miles to over 30,000 miles (see Fogel (1964) and Fishlow (1965)). Also, starting in 1846, and typically following railroad tracks, the telegraph spread across the country (see Duboff, 1980, 1983, 1984; Thompson, 1947).

Technological change, in the form of the railroad and the telegraph, eased the cost of note redemption and made information flow much faster. The reductions in travel times were dramatic. For example, between 1836 and 1862 the travel time between Philadelphia and Boston was cut by 65% (to fourteen hours) (see Gorton (1989c)).

The simple note pricing model developed here provides a framework for addressing these issues. The main result of the model is the demonstration that a bank note is equivalent to risky debt with maturity equal to the time it takes to return from the particular location of the note holder to the site of the issuing bank. In that case standard Black and Scholes (1973) option pricing theory can be used to price the bank notes. This model then provides the basis for empirical tests.

To analyze these pricing issues, I use a newly discovered complete set of bank note discounts or prices from a bank note reporter, as explained below. The data consist of monthly bank note prices of over 3,000 banks in the US and Canada traded in the Philadelphia bank note market from February 1839 to December 1858. Also necessary for the analysis, given the technological change in transportation, are time series of measures of the durations

and costs of trips from Philadelphia to the locations of these North American banks. Here, such measures are constructed from pre-Civil War travelers' guides.

The paper proceeds as follows. Section 4.2 discusses the workings of the bank note market, and introduces the data source. Section 4.3 presents an overview of the data. In Section 4.4 the note pricing model is explained. The implications of the model are confronted by the data in Section 4.5. Finally, Section 4.6 concludes.

4.2. THE BANK NOTE MARKET

Prior to the Civil War, banks issued distinct private currencies. Following the demise of the Second Bank of the United States which President Andrew Jackson refused to recharter in 1832, some states followed the lead of New York State which passed the Free Banking Act in 1838. The Act allowed anyone to open a bank, with the restriction that the private money issued by the bank be backed by designated securities deposited with state regulatory authorities.[3] Banks in chartered banking systems also were allowed to issue private money, but entry was more restricted.[4] I concentrate on the American Free Banking Era, 1838–1863, because of data availability, as described below. Hundreds of distinct private monies, called bank notes, circulated as media of exchange during the period.

Table 4.1 lists the states which adopted free banking systems and the states which did not adopt free banking, but continued as chartered banking systems. It is important to note that most states that adopted free banking did so in the 1850s. Prior to that time New York is the only example of a state which adopted free banking and which saw many new banks open.

A bank note was a small denomination noninterest-bearing, perpetual, debt obligation of the issuing bank used as a medium of exchange. The note bearer had the right to present the note for redemption at par at the issuing bank at any time.[5] Despite government enforcement of various regulations there was always

3. Free banking laws varied by state but contained some common features. Typically, banks had to back their note issuance with designated state bonds deposited with state banking authorities. Bank notes were printed and registered under the direction of state authorities. Sometimes stockholders faced double liability. Free banking was effectively ended with passage of the National Banking Acts, passed during the Civil War. Further background can be found in Dewey (1910), Hammond (1957), Grant (1857) and Cleaveland (1857).

4. Chartered banking systems were sometimes subject to abuse so that entry into banking was not always difficult. See Chaddock (1910), Hammond (1957, pp. 332–337), Knox (1903, p. 413), Ng (1987) and Sylla (1985).

5. Note holders were the senior claimants on the bank (see Breckenridge, 1899).

Table 4-1. STATES WITH AND WITHOUT FREE BANKING LAWS BY 1860

States with Free Banking Laws	Year Law Passed	States Without Free Banking Laws
Alabama	1849[b]	Arkansas
Connecticut	1852	California
Florida	1853[b]	Delaware
Georgia	1838[b]	Kentucky
Illinois	1851	Maine
Indiana	1852	Maryland
Iowa	1858[b]	Mississippi
Louisiana	1853	Missouri
Massachusetts	1851[b]	New Hampshire
Michigan	1837[a]	North Carolina
Minnesota[d]	1858	Oregon
New Jersey	1850	Rhode Island
New York	1838	South Carolina
Ohio	1851[c]	Texas
Pennsylvania	1860[b]	Virginia
Tennessee	1852[b]	
Vermont	1851[b]	
Wisconsin	1852	

[a] Michigan prohibited free banking after 1839 and then passed a new free banking law in 1857.

[b] According to Rockoff, very little free banking was done under the laws in these states.

[c] In 1845, Ohio passed a law that provided for the establishment of "Independent banks" with a bond-secured note issue.

[d] Montana became a state in 1889. The free banking law was passed by a territorial legislature.

SOURCE: Rockoff (1975, pp. 3, 125–130) as compiled by Rolnick and Weber (1983, p. 1082).

the possibility of a loss to the bearer of a bank note. The risk of bank failure, and consequent loss to note holders, varied by state for a variety of reasons other than that banks specialized in lending to borrowers with risks specific to their region. For example, bank default probabilities appear to have differed because state regulatory systems, and the degree of enforcement, varied. There was a distinction between free and chartered systems, but also variation within each type of system.

While the focus of previous research has been on the distinction between the type of banking system, free or chartered, banking systems differed in other, perhaps more important, ways. First, some banking systems allowed branching, while others did not. State bank charters limited banks' operations to

that state (for their deposit business if not their loan business). Most states also prohibited branch-banking within the state. This seems to have been unfortunate since the branch-banking states (Virginia, North Carolina, South Carolina, Georgia, and Tennessee) appear to have been less prone to panics and bank failure, possibly because of the effects of diversification admitted by branching. Also, branch systems allowed for easy interbank loans in times of emergency (see Schweikart, 1987; Calomiris, 1989; Calomiris and Schweikart, 1988).

A second dimension of state heterogeneity concerns note insurance funds. Some states sponsored insurance funds, while others did not. In general, evidence suggests that banks in states with successful mutual-guarantee or coinsurance systems (Indiana, Iowa, and Ohio) fared better than their counterparts in states without insurance. Banks covered by insurance suffered fewer failures and losses and fared better during panics. For example, in Indiana no insured bank failed during the thirty years the fund was in operation. (New York, Vermont and Michigan had less successful insurance systems.) (See Calomiris, 1989.)

A third way in which banking systems varied concerns the presence or absence of bank coalitions. The default risk associated with bank debt, in the form of bank notes, appears to have been reduced by organizations of banks which enforced their own restrictions on member bank risk-taking activity. The Suffolk system of New England is the main example of such self-regulation. The Suffolk Bank is often viewed as performing a central bank-like role in providing a clearing system for bank liabilities and concomitantly playing a regulatory role with respect to other banks.[6] By the end of the Panic of 1839, for example, only four out of 277 banks in New England outside of Rhode Island suspended convertibility of notes into specie, and they remained solvent. In other areas of the country failure rates were much higher. For example, 13.4% of the banks in Ohio, Illinois and Michigan failed.

The evidence strongly suggests that banks in branched systems, banks covered by well-run state insurance programs, and banks which were members of well-functioning bank coalitions were less prone to fail or suspend convertibility during panics. When failure did occur, banks in these systems had smaller losses. It is not known how these factors interacted with the factor which has received relatively more attention, namely, whether the system was a free or chartered banking system.

6. The Suffolk Bank system was a mechanism for clearing bank notes. Its effectiveness depended on the ability of the Suffolk Bank, a large bank at the center of the system, to control the risk-taking activities of the member banks. See Mullineaux (1987), Dewey (1910), and Whitney (1878). Gorton (1989a) presents a theoretical rationale for such bank coalitions.

4.3. BANK NOTE REPORTERS AND THE BEHAVIOR OF BANK NOTE PRICES

Once in circulation notes traded in informal secondary markets operated by note brokers. Note brokers were sometimes banks that quoted prices at which they were willing to buy and sell notes. Also, nonbank firms bought and sold notes, advertising their services in newspapers. Note brokers, often called "Exchange and Brokers' Offices," gathered information on banks, quoted bid and ask prices, often bought notes at discounts and, possibly, redeemed them at the issuing bank. Note reporters, small newspapers, reported the prices at which notes traded in the secondary markets. Agents offered unfamiliar notes consulted such publications to price the notes and determine their authenticity. Sumner (1896) explains how agents relied on bank note reporters to value notes of distant and unfamiliar banks:

> It is difficult for the modern student to realize that there were hundreds of banks whose notes circulated in any given community. The bank notes were bits of paper recognizable as a specie by shape, color, size and engraved work. Any piece of paper which had these came within the prestige of money; the only thing in the shape of money to which the people were accustomed. The person to whom one of them was offered, if unskilled in trade and banking, had little choice but to take it. A merchant turned to his 'detector.' He scrutinized the worn and dirty scrap for two or three minutes, regarding it as more probably 'good' if it was worn and dirty than if it was clean, because those features were proof of long and successful circulation. He turned it up to the light and looked through it, because it was the custom of the banks to file the notes on slender pins which made holes through them. If there were many such holes the note had been often in the bank and its genuineness ratified.

Such bank note reporters were obtained like other newspapers, by subscription or from a newsstand. Typically, the reporters were printed monthly.[7]

The data used in this study are from *Van Court's Counterfeit Detector and Bank Note List,* a bank note reporter printed in Philadelphia monthly from February 1839 through December 1858. It is a small tabloid which lists discounts on the notes of the banks of twenty–nine states and territories and three provinces of Canada. Table 4.2 lists the coverage dates and localities of the reporter. Further detail on the data is provided by Gorton (1989b).

The prices quoted by *Van Court* are not necessarily transactions prices. *Van Court* never explained exactly where the prices came from and never provided

7. See Dillistin (1949) for a discussion of bank note reporters.

volume data. But, it is not likely that every note for which *Van Court* quoted a price actually traded that month. Since the purpose of the reporter was to provide a price quotation to consumers on every conceivable note which might appear in a transaction, the coverage is extensive. Evidence suggests that the volume of notes circulating with origins outside the local area was sizeable. For example, Knox (1969, p. 368) notes that in 1857 the Suffolk Bank redeemed almost $400 million worth of other banks' notes. He also points out that for many years "Connecticut bank notes had been eagerly sought after for circulation in Ohio, Indiana and other Western States . . ." (p. 384). These observations are consistent with the sizeable inter-regional trade flows in ante bellum America. Fishlow (1964) presents quantitative evidence on these flows and Lindstrom (1975) specifically discusses Philadelphia.

Not all banks issuing private money during the Free Banking Era are covered by *Van Court*. Comparing Table 4.1 to Table 4.2, note that Oregon, Texas, California, and Minnesota were not covered by *Van Court*. Bank notes from these locations, if listed by *Van Court*, were described as of "uncertain" value. Also, only partial coverage is provided for many locations, such as Canada, Wisconsin, and Montana. It is noteworthy that the locations which are not covered, or for which coverage is partial, are typically locations long distances from Philadelphia. While this is consistent with the notion that distance from Philadelphia back to the issuing bank is important in note pricing, it also suggests that the situation is more complicated. For example, Montana is further away than Minnesota. Yet, Minnesota, generally considered to be an example of a failed free banking state, is never covered. Below these observations about distance will be made more precise.

4.3.1. Free Banking States, Chartered Banking States

Tables 4.3 and 4.4 provide summaries of the data from *Van Court* for two states. The two states, to some extent representative of the variety of state experiences, are Indiana and North Carolina. (Gorton (1989a) contains similar tables for all other locations.) Indiana adopted free banking in 1852. North Carolina was a chartered banking state for the entire period.

The tables list a variety of information about the note discounts, including the "average modal discount" which is the annual average of the monthly modes. At each monthly date the bank notes of most banks at each particular distant location are trading at the same discount in Philadelphia. This number is the modal discount for the month. The annual average of the monthly modal discounts is the "average modal discount." The column entitled the "average modal percent" gives the average of the monthly percentages of the total number of banks in that location which had the modal discount. The mean discount is higher than the modal discount because many of the banks with discounts listed

Table 4-2. COVERAGE OF VAN COURT'S BANK NOTE REPORTER: STATES AND DATES

States with Complete Coverage, February 1839–December 1858		**States with Incomplete Coverage**[c]		**States Listed as "Uncertain" or Not Listed**
United States	**Canada**	**United States**	**Canada**	
Alabama	Canada[b]	Arkansas (1840–58)	New Brunswick (1840–48)	Iowa territory
Connecticut	Nova Scotia	Florida (1842–58)		Minnesota
Delaware		Illinois (July 1856–58)		Missouri
District of Columbia		Indiana (1857)		Texas
Georgia		Michigan (1853)		
Kentucky		Mississippi (1839, 1841–43, 1852–58)		
Louisiana		Nebraska (1840–47)		
Maine		New Hampshire (1857–58)		
Maryland		Virginia (1846–47, 1853–54)		
Massachusetts		Wisconsin (1839–55)		
Montana[a]				
Pennsylvania				
New Jersey				
New York				
North Carolina				
Ohio				
Rhode Island				
South Carolina				
Tennessee				
Vermont				

[a] Montana became the 41st state in 1889.
[b] Canada includes banks located in provinces other than Nova Scotia or New Brunswick.
[c] Incomplete coverage means that the *Van Court Bank Note Reporter* did not quote a price for banks in that state that month. The state may have been listed, though, and the notes of banks in that state described as "all uncertain." Dates in parentheses indicate periods for which the data was missing.

by *Van Court* are insolvent.[8] The tables also provide the number of banks in existence each year. The leverage measures, constructed from the 1876 Comptroller of the Currency *Annual Report,* are measures of the annual aggregate leverage of banks in the particular location.

8. The notes of insolvent banks had positive prices because insolvent banks were liquidated over a period of time. During the liquidation period some notes were redeemed and the value of the remaining assets fluctuated. Rockoff (1974a,b) also makes this point. Van Court does not indicate whether a bank is insolvent or not.

Table 4-3. SUMMARY OF INDIANA BANK NOTE DISCOUNT DATA[a]

Year	(1) Mean Discount	(2) Standard Deviation	(3) Minimum Discount	(4) Maximum Discount	(5) Average Mode[b]	(6) Annual Standard Deviation of Mode[c]	(7) Average Modal%[d]	(8) Number of Banks[e]	(9) Notes Total assets	(10) Notes +Deposit Total assets	(11) Specie Total assets
1839	4.36	0.861	3.250	5.500	4.364	0.861	100.00	1	0.024	0.101	0.210
1840	4.83	0.389	4.000	5.000	4.833	0.389	100.00	1	0.026	0.083	0.156
1841	7.41	1.062	5.000	9.000	7.417	1.062	100.00	1	0.026	0.099	0.166
1842	21.67	23.800	5.000	70.000	10.417	4.940	87.50	2	0.027	0.076	0.187
1843	19.61	19.680	2.000	60.000	2.773	0.984	50.00	2	0.007	0.048	0.166
1844	10.01	8.640	1.500	22.500	1.688	0.155	50.00	2	0.012	0.055	0.209
1845	9.75	7.953	1.750	17.500	2.000	0.204	50.00	2	0.025	0.080	0.209
1846	7.31	6.422	1.500	17.500	2.125	0.506	50.00	2	0.014	0.082	0.183
1847	4.42	3.151	1.250	7.500	1.333	0.123	50.00	2	0.019	0.088	0.156
1848	4.81	2.762	1.750	7.500	2.125	0.433	50.00	2	0.043	0.137	0.156
1849	4.55	3.016	1.250	7.500	1.604	0.249	50.00	2	0.022	0.106	0.192
1850	4.48	3.089	1.000	7.500	1.458	0.209	50.00	2	0.015	0.111	0.187
1851	4.91	4.478	1.000	20.000	1.271	0.250	50.00	2	0.028	0.028	0.031
1852	9.05	9.363	0.750	20.000	1.313	0.188	48.10	5	0.043	0.129	0.161
1853	1.58	2.065	0.500	20.000	1.230	0.072	80.92	22	0.067	0.173	0.160
1854	6.60	6.251	1.130	15.000	5.105	6.012	99.72	91	0.046	0.159	0.125
1855	19.24	11.130	1.000	50.000	20.667	12.280	51.02	110	0.042	0.147	0.095
1856	26.73	27.980	1.000	80.000	5.000	0.000	33.24	97	0.046	0.198	0.153
1857	-	-	-	-	-	-	-	-	0.044	0.189	0.117
1858	9.70	17.04	5.000	75.000	5.000	-	90.91	33	0.043	1.197	0.163

[a] The missing values do not mean that the bank note reporter did not report the data. Rather, the reporter would list all the bank notes of the state as "uncertain."

[b] The average mode is the annual average of the twelve monthly modal discounts.

[c] The annual standard deviation of the mode measures the variation of the monthly modal discounts during the year.

[d] The modal percentage is the percentage of total banks with modal discounts. The average modal percentage is the annual average of the twelve monthly modal percentages.

[e] The number of banks in existence during the year.

Table 4-4. SUMMARY OF NORTH CAROLINA BANK NOTE DISCOUNT DATA[a]

Year	(1) Mean Discount	(2) Standard Deviation	(3) Minimum Discount	(4) Maximum Discount	(5) Average Mode[b]	(6) Annual Standard Deviation of Mode[c]	(7) Average Modal%[d]	(8) Number of Banks[e]	(9) Notes Total assets	(10) Notes +Deposit Total assets	(11) Specie Total assets
1839	3.24	0.95	2.000	5.000	3.188	0.98	100.00	3	0.021	0.114	0.114
1840	1.88	0.66	1.000	3.000	1.875	0.68	100.00	3	0.035	0.116	0.091
1841	2.33	0.76	1.500	4.000	2.333	0.78	100.00	3	0.036	0.115	0.130
1842	3.96	2.30	2.000	10.000	8.958	2.37	100.00	3	0.054	0.147	0.143
1843	1.88	0.30	1.500	2.500	1.875	0.31	100.00	3	0.045	0.136	0.160
1844	1.27	0.07	1.250	1.500	1.271	0.07	100.00	3	0.035	0.133	0.152
1845	1.46	0.14	1.250	1.750	1.458	0.14	100.00	3	0.047	0.139	0.166
1846	1.78	0.22	1.500	2.250	1.729	0.23	100.00	3	0.061	0.150	0.176
1847	1.40	0.19	1.250	1.750	1.396	0.20	100.00	3	0.046	0.136	0.190
1848	2.08	0.38	1.750	2.750	2.083	0.39	100.00	4	0.039	0.118	0.176
1849	1.73	0.26	1.500	2.250	1.729	0.27	100.00	4	0.037	0.106	0.182
1850	1.35	0.12	1.250	1.500	1.354	0.13	100.00	4	0.050	0.139	0.176
1851	1.38	0.19	1.250	1.750	1.375	0.20	100.00	5	0.051	0.149	0.172
1852	1.34	0.17	1.000	1.500	1.344	0.18	100.00	7	0.051	0.149	0.172
1853	1.00	0.00	1.000	1.000	1.000	0.00	100.00	6	0.037	0.141	0.145
1854	2.64	2.75	1.000	15.000	1.796	0.68	81.98	11	0.043	0.164	0.129
1855	1.95	0.54	1.500	3.000	1.958	0.56	100.00	13	0.029	0.109	0.094
1856	1.38	0.13	1.250	1.500	1.375	0.13	100.00	13	0.022	0.099	0.095
1857	2.70	3.56	1.000	30.000	2.500	2.76	98.08	13	0.024	0.100	0.076
1858	3.43	4.13	1.000	30.000	2.458	1.77	91.78	13	0.026	1.098	0.072

[a] The missing values do not mean that the bank note reporter did not report the data. Rather, the reporter would list all the bank notes of the state as "uncertain."
[b] The average mode is the annual average of the twelve monthly modal discounts.
[c] The annual standard deviation of the mode measures the variation of the monthly modal discounts during the year.
[d] The modal percentage is the percentage of total banks with discounts. The average modal percentage is the annual average of the twelve monthly modal percentages.
[e] The number of banks in existence during the year.

Indiana is often viewed as one of the worst examples of free banking, though its insurance system is considered to have been a success. Between 1834 and 1853 the State Bank of Indiana was the only bank in the state. It had branches throughout the state, but the "branches" were separately owned and operated. The bank easily weathered the storm of the Panic of 1837. In 1853, however, the state constitution was changed to allow free banking. (Free banks were not covered by insurance.) As can be seen in Table 4.3, the number of banks quickly increased. The modal discount also increased dramatically. The modal percentage falls by one half implying that the newly entering banks' notes were more heavily discounted.

During the Panic of 1857 two thirds of the Indiana banks went bankrupt. In Table 4.3 there is no entry for this year because *Van Court* listed Indiana banks as all uncertain (even before the panic). Rockoff (1974b) cites evidence suggesting that the problem in Indiana was that the state auditor may have valued Indiana bonds, used to back bank note issues, at par when their market value was less than par.[9]

North Carolina is an example of a chartered banking system (without an insurance system). North Carolina authorized an official state bank in 1854. This bank had branches in four cities and agencies in six others, but did not have a monopoly because the legislature also authorized two other banks. The state government appears to have overseen these banks carefully. Between 1847 and 1860 the state authorized the incorporation of fourteen new private banks with twenty–six branches. These new banks were allowed to receive deposits but could not "issue any bill, note or other device in the nature of a bank note" (see Knox, 1969). Notably, as shown in Table 4.5, both the modal discount and the standard deviation of the modal discount are low compared to the free banking states.[10]

In Tables 4.3 and 4.4 the modal discount is most relevant. The modal discount is the focus of the subsequent empirical work because it represents the price at which the notes of solvent banks traded. In the Philadelphia note market, the notes of most of the banks at any specific distant location traded at the same price, the modal discount. Below I provide a theoretical reason for this phenomenon. All other discounts of banks at the particular location are higher, suggesting that those banks were insolvent. (In fact, for a sample of New York banks, I verified that banks trading at the higher nonmodal discount are insolvent. Insolvent banks were liquidated over a period of time during which their notes continued to trade.)

9. For a further discussion of Indiana see Harding (1895) and Dewey (1910). See Calomiris (1989) on Indiana's insurance system.

10. For more information on North Carolina see Schweikart (1987).

Several other important observations can be made about Tables 4.3 and 4.4. For any given location, the modal discount varies substantially over time and does not decline smoothly as might be predicted from a simple notion of how the discount relates to the diffusion of the railroad and the telegraph. Not only does the discount not decline smoothly, but the effects of the introduction of the railroad and the telegraph are not obvious. It seems clear that the modal discount is not solely a function of distance from Philadelphia to the issuing bank, though more will be said about this below. Finally, note the variation in the modal percentage over time for a given location. This reflects the number of insolvent banks with notes still in circulation.

4.3.2. Note Discounts, Railroads, and Panic

Table 4.5 provides a summary of the data from *Van Court* for the year 1839, the beginning of the sample period. The table shows the monthly modal discounts for each location on which *Van Court* reported in each of that year. During this period there was a banking panic, visible in Table 4.5 as negative discounts.[11]

As expected, the modal discount for Pennsylvania is always zero. Also, the modal discounts for New England states tend to be lower than other states, possibly reflecting the Suffolk system. But another possibility is simply that New England was a long-settled, possibly less risky, region. Moreover, there was almost no free banking in New England. But, it has been argued that state legislatures in this region were quick to grant bank charters so that entry into banking was similar to a free banking state (see Sylla, 1985).

Table 4.5 also makes clear that distance is not related to note discounts in any simple way. The table provides several examples where the discounts are *higher* on the notes of banks at locations which are *closer* to Philadelphia. For example, the discounts on the notes of Tennessee are zero in Table 4.5. Yet, Tennessee is clearly farther from Philadelphia than many of the other locations. The Tennessee banking system was dominated by an official state bank, the Bank of Tennessee, which at the beginning of the period was fully backed by the state

11. In Table 4.7 the reader will notice that there are some negative entries for modal discounts. These occur during the Panic of 1839 (and during a few months of the Panic of 1857). During periods of suspension of convertibility following banking panics *Van Court* apparently switched from quoting prices in terms of gold to quoting prices in terms of Philadelphia bank notes. During a period of suspension it was not possible to convert bank notes into specie on demand. Apparently, for this reason, *Van Court* switched to quoting prices in terms of Philadelphia bank notes during suspensions. Thus, in terms of Philadelphia notes, the notes of some banks would be worth a premium though still at a discount in terms of gold. See Gorton (1989b) for details. On the Panic of 1857 see Van Vleck (1943).

Table 4-5. SUMMARY OF 1839 DISCOUNT DATA

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sept.	Oct	Nov.	Dec.
1) Alabama	—	3.50	3.50	10.00	10.00	10.00	14.00	12.50	15.00	12.50	10.00	2.00
2) Arkansas	—	12.50	15.00	15.00	15.00	15.00	15.00	15.00	—	—	—	—
3) Connecticut	—	0.75	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	−3.00	−5.00
4) Delaware	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5) Washington, DC	—	0.50	0.50	0.88	0.88	0.88	0.88	0.88	0.88	1.50	2.00	1.00
6) Georgia	—	3.50	3.75	5.50	5.50	5.50	5.50	4.50	5.00	10.00	10.00	5.00
7) Illinois	—	3.25	3.25	4.00	4.00	4.00	4.00	4.00	5.50	5.50	6.50	6.50
8) Louisiana	—	1.25	1.25	3.50	3.50	3.50	3.50	5.00	7.00	7.00	6.00	0.00
9) Maine	—	0.75	0.75	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	−5.00
10) Massachusetts	—	0.75	0.75	1.00	1.00	1.00	1.00	1.00	1.00	−3.00	−3.00	−5.00
11) Michigan	—	1.50	10.00	10.00	10.00	10.00	10.00	8.00	5.00	−3.00	7.00	7.00
12) Montana	—	4.00	4.00	4.00	4.00	4.00	4.50	4.50	6.00	7.00	7.00	5.00
13) Maryland	—	0.50	0.50	1.00	0.50	0.375	0.375	0.50	0.50	2.00	0.75	0.75
14) North Carolina	—	2.50	2.50	3.00	3.00	3.00	3.00	3.00	4.00	5.00	5.00	2.00
15) Nebraska	—	20.00	20.00	20.00	20.00	20.00	20.00	20.00	20.00	0.00	—	—
16) New Hampshire	—	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	−5.00
17) New Jersey	—	0.75	0.75	1.00	1.00	1.00	1.00	1.00	1.00	1.00	−3.00	−5.00
18) New York	—	0.75	0.75	1.00	1.00	1.00	0.75	1.00	1.00	−6.00	−5.00	−5.00
19) Ohio	—	3.25	3.25	4.00	4.50	4.00	4.00	4.50	5.00	5.50	5.50	5.00
20) Pennsylvania	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
21) Rhode Island	—	0.75	0.75	1.00	1.00	1.00	0.75	1.00	1.00	−3.00	0.00	0.00
22) South Carolina	—	2.50	2.50	3.00	3.00	2.75	2.75	3.00	5.00	7.00	3.00	0.00
23) Tennessee	—	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
24) Vermont	—	0.75	0.75	1.00	1.00	1.00	1.00	1.00	1.00	−2.00	−3.00	−5.00
25) Virginia	—	0.75	0.75	1.00	1.25	1.25	1.00	1.00	1.50	4.00	2.00	1.00

and acted like a central bank (see Campbell, 1932). Also, note that the discounts of Vermont's banks' notes are the same as those of New Jersey bank notes. There are many examples of this sort in the data, though New Jersey borders Pennsylvania.

4.3.3. Travelling from Philadelphia to the Bank of Issuance

In order to exercise the redemption option feature of the note contract, the note bearer had to travel to the location of the issuing bank. Also, for much of the period and many locations, information would have to have travelled by the same mode of transportation that people used. Consequently, the cost of such a trip in terms of time or money would naturally seem to be related to the note discounts or prices. Banks which are more distant from Philadelphia should have notes which are more heavily discounted, *ceteris paribus*. In fact, a traditional hypothesis explaining the cross-section variation in note discounts is that the cost of returning from the note holder's location to the bank of issuance is the dominant factor. Since banks were risky institutions it is not clear to what extent the discounts reflect travel costs and to what extent they reflect other factors.

In order to analyze the relations between travel costs and note discounts, and to evaluate the note pricing model to be described in Section 4.4, measures of the distance from Philadelphia back to the location of the banks covered by *Van Court* are needed. In particular, measures of the costs and the durations of such trips are needed. Such measures would capture the dramatic diffusion of the railroad across the eastern part of the US, as well as the improvements in canals and steamships.

Gorton (1989c) constructs transportation costs and trip duration indices using pre-Civil War travelers' guides and historical information on the costs and speeds of various modes of travel. The travellers' guides provided the pre-Civil War traveler with the most commonly used routes from Philadelphia to various other locations in North America. The guides detail the route to be taken, and indicate whether each leg of the journey was to be by stagecoach, canal, steamboat, or railroad. Combining this information with estimates of the speeds and costs of each mode of transportation, indices were constructed for three years: 1836, 1849, and 1862 (the only years for which such guides could be located).

Examination of these indices confirms that improvements in transportation technology were dramatic. The time and costs of a trip from Philadelphia to other locations in North America were greatly reduced. Figure 4.1 graphically portrays the reductions in the durations of trips from Philadelphia to the capitals of selected other locations.

To what extent does the distance to the issuing bank explain cross-section variation in the discounts? Table 4.6 reports the (Spearman rank) correlations of

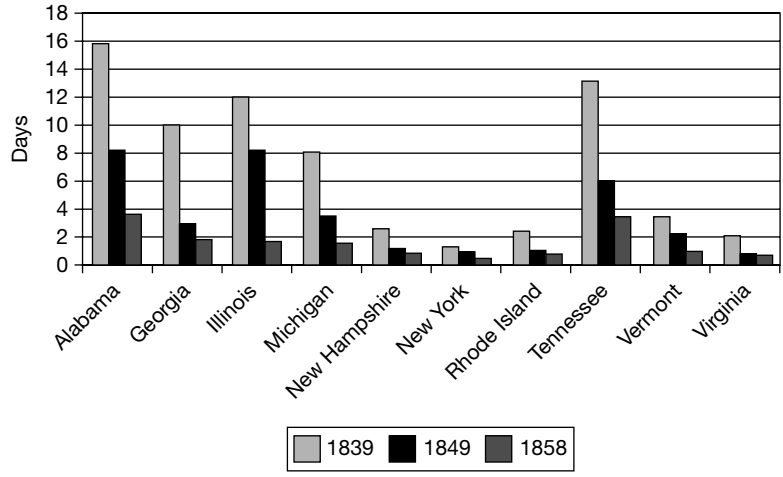


Figure 4.1 Trip times from Philadelphia to ten state capitals

discounts with the measures of the cost of the return trip and the duration of the return trip.[12] Cross-section regressions of the (annual average) modal discount on both transportation indices jointly yield:

For 1839:

$$\text{Modal discount} = -1.07 - 0.44 * \text{Trip Cost} + 0.122 * \text{Trip Time},$$
$$(4.3) \quad (4.2) \qquad\qquad (5.3)$$
$$R^2 = 0.31.$$

For 1849:

$$\text{Modal discount} = 0.326 - 0.011 * \text{Trip Cost} + 0.04 * \text{Trip Time},$$
$$(1.19) \quad (0.27) \qquad\qquad (3.05)$$
$$R^2 = 0.12.$$

For 1858:

$$\text{Modal discount} = 0.333 - 0.059 * \text{Trip Cost} + 0.067 * \text{Trip Time},$$
$$(3.3) \quad (4.08) \qquad\qquad (7.3)$$
$$R^2 = 0.11.$$

12. Note that only the year 1849 is the correct match of the distance data with the discount data. Unfortunately, the distance data for 1836 had to be matched with 1839. Similarly, 1858 and 1862 were matched.

Table 4-6. Correlations Between Discounts and Distance[a]

	Cost of Trip	Trip Duration	Modal Discount	Avg. Nonmodal Discount
		1839		
Cost of trip	1.000	0.96	0.656	0.525
	(0.000)	(0.000)	(0.001)	(0.021)
Trip duration		1.000	0.653	0.523
		(0.000)	(0.001)	(0.022)
Modal discount			1.000	0.593
			(0.000)	(0.008)
Avg. nonmodal discount				1.000
				(0.000)
		1849		
Cost of trip	1.000	0.95	0.794	0.280
	(0.000)	(0.000)	(0.000)	(0.261)
Trip duration		1.000	0.787	0.300
		(0.000)	(0.001)	(0.226)
Modal discount			1.000	0.422
			(0.000)	(0.081)
Avg. nonmodal discount				1.000
				(0.000)
		1858		
Cost of trip	1.000	0.96	0.800	0.674
	(0.000)	(0.000)	(0.000)	(0.003)
Trip duration		1.000	0.789	0.669
		(0.000)	(0.001)	(0.003)
Modal discount			1.000	0.317
			(0.000)	(0.215)
Avg. nonmodal discount				1.000
				(0.000)

[a]Pearson correlation coefficients. Probability of zero correlation in parentheses. 288 observations for each year. See Gorton (1989d) for details.

t-statistics are given in parentheses. The results in Table 4.6 and the above regressions confirm the popular notion that the return trip to the issuing bank is a prime determinant of the discount in cross-section. The traditional hypothesis does fairly well.

But travel time by itself does not appear to be a completely satisfactory explanation. The main difficulty concerns examples like those noted above where discounts were higher on the notes of banks which were relatively closer to Philadelphia. Either there are other important determinants of the discounts or the note market was inefficient. Are these other determinants the risk attributes

of the banking system of that state? Were these risks priced? In order to analyze this question the next section presents a model of bank note pricing.[13]

4.4. PRICING BANK NOTES

In this section a very simple, stylized, model of bank note pricing is presented. The model is based on Svensson (1985). (See also Gorton, 1996.) The goal of the model is to relate the note price to the duration of a trip back to the bank of issuance. Then the above transportation indices can be used to study the effects of technological change.

4.4.1. A Model of Bank Note Pricing

Assume that agents are spatially separated. Let "d" be a measure of the distance from an agent's home to the market which is the location of the agent's trade at time t. Thus, d indexes location. (A time subscript on d will be omitted, except as necessary.) Each agent owns a firm at the home location. Firms at each location receive a stochastic endowment of a single nonstorable good, $y(d)_t$. Output is assumed to be independently, identically, lognormally distributed at each date t and location d. The standard deviation of output at location d is given by $\sigma(d)$.

Each household-firm begins period t with equity, Q_{t-1}, and debt, D_{t-1}, outstanding. These are claims on the household's endowment stream. The debt of a firm consists of small denomination noninterest-bearing perpetuities with embedded American put options allowing conversion of the debt into consumption goods on demand at par. The debt is called "bank notes." All output not used to honor debt is paid out as dividends since goods are nonstorable. Each household is a money-issuing firm so the terms "bank," "household," and "firm" all refer to the same economic unit.

The representative household (at a representative location) is assumed to prefer goods procured from locations further from home rather than procured nearer home:

$$E_t \left\{ \sum_{j=t}^{\infty} \beta^{j-t} U(C, d) \right\}, \tag{4.1}$$

13. It is worth noting that a linear Tobit model with the modal discount as dependent variable and trip time and risk measures as independent variables does very poorly (see earlier versions of this paper). If the model of the next section is correct, then there is a nonlinear relationship between note price and the other variables, suggesting that the linear specification is incorrect.

where $0 < \beta < 1$, $U'_C > 0$, $U''_\alpha < 0$, $U'_d > 0$, $U''_{dd} < 0$. The assumption that utility depends on distance is intended to capture the notion of a division of labor. The introduction of distance as an argument of the utility function is a device to model a desire for goods from other locations.[14]

Each household is to be thought of as consisting of a buyer and a seller, as in Lucas (1980). The seller stays at home and sells the output of the firm receiving bank notes in exchange. The buyer chooses to travel a distance, *d,* to buy consumption goods, paying for them with bank notes. (Assume the buyer chooses the distance *d,* but that the direction is random. All expectations below will be taken over this uncertainty.) Only one market can be visited at each date *t.* Buyers face a cash-in-advance constraint which can only be satisfied by bank notes.[15] Let $P_t(d)$ be the price (in terms of consumption units) of bank notes issued by the representative agent and traded at location, *d,* at time *t.* Thus, the buyer is constrained by

$$C_t \leq \sum\nolimits_d P_t(d) D_{t-1}(d), \tag{4.2}$$

which is the cash-in-advance constraint. In Eq. (4.2), the buyer carries a portfolio of bank notes from banks at different distances (indicated by the argument d) from the market that is chosen for transactions at date *t.* This market will be a distance d away from the home location.

The sequence of events in a period, *t,* is as follows. At the start of period *t,* the current state, $y(d)_t$, is learned for each location, d.[16] Then the goods market opens. The household buyer travels the distance d carrying the predetermined portfolio of bank notes. (The portfolio was held over from date $t - 1$.) The buyer purchases C_t consumption units from sellers at location

14. In ante bellum America there was a spatial division of labor. The traditional thesis concerning this division of labor was articulated by Schmidt (1939) and Callender (1909). Also see Mercer (1982) and Pred (1980). The main point is that interregional trade flows between different locations were sizeable. It is not known to what extent these flows imply a large volume of bank notes moving around the country.

15. For simplicity the model omits specie as an alternative medium for satisfying the cash-in-advance constraint. Since, as will be seen, a capital loss is associated with carrying notes to distant markets, gold or silver would appear to be preferable as a means of exchange. Thus, unless there is some cost to using gold or silver, bank notes would not circulate much beyond the location of the bank of issuance. During the ante bellum period the costs of using specie were sizeable. First, specie is heavy and difficult to transport. Second, insofar as there were coins available, there was a confusing array of denominations because many (possibly most) of the coins in circulation were foreign. The US mint was incapable of reminting the foreign coins because of poor mechanical minting equipment and because of the transportation costs of moving specie. See Carothers (1930) and Dewey (1910). Third, there was a shortage of small coins which was met by bank notes (see Carothers, 1930, p. 79).

16. This assumption is consistent with the existence of the telegraph.

d, using bank notes, and then returns home. Meanwhile, the seller sells goods in the home market, receiving bank notes in exchange for consumptions goods. After the goods market closes, and buyers have returned home, the securities market in which notes and shares are traded opens at each location. At this time a household chooses a portfolio of notes and shares and, in particular, may decide to redeem some notes. When the securities market opens, prices for the notes will already have been established in the goods market. At those prices households decide to hold notes or redeem notes, depending on whether they expect to travel a greater or lesser distance next period.

In order to model the idea that note redemption requires a time consuming trip, the following assumption is made. The receipt of a note issued by a firm at distance d from the issuer's location is assumed to imply that it takes d periods to return for redemption, if the holder wants to redeem it. In other words, there is assumed to be an asymmetry between household buyers and sellers. Buyers can carry a note a distance d during a single period, but, a seller who receives the note requires d periods to return it if the redemption option is exercised. Thus, it is costly to redeem notes in the sense that it is time consuming. Since it is time consuming to redeem notes, the amount of debt which will actually be redeemed in period t was, in fact, determined at past dates, and so is predetermined at the start of period t.

The amount of debt that will be redeemed in the current period depends upon the profile of locations, and hence, dates in the past, from which debt was sent for redemption. Notes sent for redemption at date t will be *in transit* for d_t periods.[17] Suppose that a note of a bank located at a distance d from the home location was sent for redemption k periods ago. This note will be in transit for d periods before it is redeemed. At any time t, if $d > k$, then the note will be redeemed in $d - k$ periods. If, at time t, $d = k$, then the note is presented for redemption in the current period. Let $D_t^R(d)$ be the amount of notes sent for redemption d periods ago. When $d = 0$ the amount of notes the bank must redeem is $D_t^R(0)$.

The situation of the firm, at time t, is as follows. When selling output at time t, the firm receives bank notes which are the obligations of banks various distances away. At the home location the amount received from sales in period t is: $\sum_d P_t(d)D_t = y_t(0)$. At the firm's own location the price of a dollar of its own notes is $P_t(0)$. This is the price at which its notes will be redeemed in period t. The amount of debt which the firm will redeem (in consumption units) is: $P_t(0)\,D_t^R(0)$. ($P_t(0) = 1$ if the firm is solvent.)

17. Once notes have been sent for redemption, it is assumed that they cannot be called back.

The firm may also issue new debt and new equity. For simplicity assume that no new equity is issued and that the face value of new debt issued, $D_t^N(0)$, always equals the face value of the amount redeemed, so long as the firm is solvent. Thus, the firm's leverage is constant. Since debt does not pay interest, the dividends the representative household pays out are always y_t.

Let $q_t(d)$ be the price of shares of banks at location d in period t and let $v_t(d)$ be the dividend paid to an owner of a share of stock issued by a household from location d. Then, the resources available to the household consist of: (i) shares and dividends, $\sum_d [q_t(d) + v_t(d)]\, Q_{t-1}(d)$; (ii) the value of the debt of other firms redeemed, $P_t(0)\, D_t^R(0)$; and (iii), any monies not spent satisfying the cash-in-advance constraint. In the securities market these resources will be used to finance: (i) a portfolio of bank shares; (ii) a portfolio of bank notes of various types to be held until the next period to finance consumption; and (iii), an amount of each bank's notes to be sent for redemption. So the budget constraint is

$$\sum_d \{q_t(d)Q_t(d) + P_t(d)[D_t(d) + D_t^R(d)]\} + P_t(0)D_t^N(0)$$
$$\leq \sum_d P_t(d)D_{t-1}(d) - C_t + P_t(0)D_t^R(0) + \sum_d [q_t(d) + v_t(d)]Q_{t-1}(d). \quad \textbf{(4.3)}$$

4.4.2. Equilibrium

The representative agent chooses a distance to travel in period t, d_t, an amount of notes of each type, d, to be sent for redemption $D_t^R(d)$, an amount of notes of each type, $D_t(d)$, to be carried to next period, and an amount of equity shares of each type, $Q_t(d)$, to hold to maximize (4.1) subject to (4.2) and (4.3). Let μ be the Lagrange multiplier associated with the cash-in-advance constraint, (i). The first-order conditions with respect to choice of $D_t(d)$, $D_t^R(d)$, d_t and $Q_t(d)$, respectively, assuming an interior solution, can be written as

$$U'_{Ct} = \beta E_t\{U'_{Ct+1}[P_{t+1}(d)/P_t(d)]\} + \beta E_t\{\mu_{t+1}[P_{t+1}(d)/P_t(d)]\}, \quad \textbf{(4.4)}$$

$$U'_{Ct}P_t(d) = \beta^d E_t\{U'_{Ct+d}P_{t+d}(0)\}, \quad \textbf{(4.5)}$$

$$U'_{dt} = -U_{dt}\sum_d P'_{dt}\{D_{t-1}(d) - [D_t(d) + D_t^R(d)]\}$$
$$+ \mu_t \sum_d P'_{dt}D_{t-1}(d), \quad \textbf{(4.6)}$$

$$U'_{Ct}q_t(d) = \beta E_t\{U'_{Ct+1}q_{t+1}(d)\}, \quad \textbf{(4.7)}$$

where E_t indicates the expectation conditional on information available at time t. (There are also transversality conditions for each note.)

Equilibrium requires that: (i) the goods market at each location clear, i.e., $C_t(d) = y_t(d)$ for each d; (ii) the market for each bank's equity clear, $Q_{t-1} = Q_{t+1}(d) = 1$, for each d; (iii) the market for each bank's debt clear, $D_{t-1}(d) = D_t^R(d) + D_t(d)$, for each d; (iv) $\sum_d v_t(d) = y_t(d)$, for each d, that is, each household pays out dividends in the amount of the firm's proceeds that period; (iv) by assumption, $D_t^R(0) = D_t^N(0)$, that is, the amount of new notes issued equals the amount retired.

The first-order condition (4.4) determines the optimal choice of $D_t(d)$, the face value amount of bank notes from location d to be carried over to next period to provide the household buyer with bank notes to satisfy the cash-in-advance constraint. A bank note dollar held to next period has a direct return, as part of wealth, the first term on the right-hand side of Eq. (4.4), and a future benefit in the form of future liquidity services when the note dollar is used to satisfy next period's cash-in-advance constraint, the second term. See Svensson (1985) for a discussion.

Conditions (4.5) and (4.7) price the firm-bank's debt and equity, respectively. Write Eq. (4.5) as

$$P_t(d) = \beta^d E_t \left\{ P_{t+d}(0) \left[U'_{Ct+d} / U'_{Ct} \right] \right\}, \tag{4.8}$$

where $P_{t+d}(0)$ is the redemption value of a note d periods from now. This price assumes a first-come-first-served rule since at date $t + d$, $D_{t+d}^R(0)$ notes have been presented for redemption, and only this debt must be honored at that time. Bankruptcy is defined by whether or not the bank can honor the amount of debt being presented for redemption, $D_t^R(0)$, and not by the outstanding amount of debt.

In considering redemption a complication arises because notes may have been sent for redemption in the past which have not yet reached the issuing bank. These notes are in transit to the bank. Suppose, for the moment, that there are no notes in transit. (This would be known at time t.) If there are no notes in transit, then there is no question of the bank defaulting prior to presentation of the notes currently being sent for redemption. The value of the bank at time t and location d is $V_t(d) = P_t(d) D_t + q_t(d) Q_t$.

We now turn to pricing the bank notes. To begin with, see Proposition 1.

PROPOSITION 1. The bank notes of a bank a distance d away are valued as risky pure discount debt claims with a maturity of d periods.

To see this note that from Eq. (4.5), which can be solved for the price of the bank note at location d, $P_t(d)$, the representative agent must, in equilibrium, be indifferent between holding a one dollar note and sending the note for redemption. The value of a note sent for redemption as is given by Eq. (4.8) values the

note as a risky debt claim maturing d periods later. Even though the debt is perpetual, from the point of view of the representative agent, since it takes d periods to redeem, it can be priced as debt of maturity d. Thus, we can state the second proposition.

PROPOSITION 2. Assume that preferences display constant relative risk aversion. Then, if $D_t^R(d)$ is the face value of the amount of debt sent for redemption at date t, from location d, its value at date t is given by

$$P_t(d) = \left[D_t^R(d)\right]^{-1}\{V_t(d)\left[1 - N(h_D + \sigma)\right] + (1 + r_f)^{-1} D_t^R(d) N(h_D)\}, \tag{4.9}$$

where $h_D, \left\{\ln\left[V_t(d)/D_t^R(d)\right] + \ln\left(1 + r_f\right)\right\}/\sigma - \sigma/2$.

σ is the standard deviation of one plus the rate of change of the value of the bank (i.e., the standard deviation of output), and r_f is the risk free rate of interest (assumed constant). N(•) indicates the cumulative Normal distribution function.[18] The proposition says that bank notes can be priced using Black and Scholes (1973) option pricing formula. The proof of this proposition is standard and due to Rubinstein (1976).

Propositions 1 and 2 were derived under the assumption that there were no notes in transit. If there are notes in transit, then, between the current date, t, and date $t + d$, these notes will, successively, be presented for redemption. These notes are more senior claimants in a sense. The bank may default on one of these payments. From the point of view of the household/bank these successive redemptions are akin to coupon payments. The stock is then a compound option because until the current amount, $D_t^R(d)$ has been redeemed at date $t + d$, the stockholders have the option of buying the option to redeem the next amount which will be presented. Under these conditions a proposition analogous to Proposition 2 can be proven. That is, assuming that preferences display constant relative risk aversion, the bank notes can be priced according to Geske's (1977) extension of Black-Scholes.

Equilibrium in the goods market requires that the note price, $P_t(d)$, adjust to clear the market given choice of location d. Then, in the securities market, notes will be demanded for satisfying future liquidity constraints. We can now inquire as to when the redemption option is worth exercising. A note dollar held must satisfy Eq. (4.4); a note dollar sent for redemption must satisfy Eq. (4.5). Thus, the option is "in the money" when a note dollar is more valuable being sent for redemption, i.e., when the value of a note given by the right-hand side of Eq. (4.5) is greater than the left-hand side and vice versa for Eq. (4.8).

18. For simplicity the model has no riskless security. However, the shadow price of a riskless bond can always be calculated. A riskless security could easily be incorporated.

4.4.3. Equilibrium Note Price Characteristics

Since bank notes can be priced using Proposition 2, Black and Scholes' option formula, some useful comparative statics are immediate.[19] In particular, the value of the notes, $P_t(d)$, varies inversely with d, σ, and leverage of the bank (see Merton, 1974). These results, will provide the basis for confronting the data, starting in the next section.

An important feature of the data is that *Van Court* quoted "all uncertain" for banks a long distance from Philadelphia, suggesting that the notes of these banks were very highly discounted, perhaps to zero. Locations even further away were not listed. The above valuation model implies that, at the same distance from the issuing bank, not all notes will circulate. Condition (4.6) determines the optimal choice of distance from home, d_t^*, the buyer should travel to buy consumption goods. To understand Eq. (4.6), recall that in equilibrium $D_{t-1}(d) = D_t(d) + D_t^R(0)$, i.e., the stock of bank notes outstanding for each bank and carried over into period t, must be divided into an amount held until next period and an amount sent for redemption.[20] Thus, in equilibrium, Eq. (4.6) becomes:

$$U'_{dt} = -\mu_t \Sigma_d P'_{dt} D_t(d). \qquad \textbf{(4.10)}$$

By Proposition 2, $P'_{dt} < 0$, i.e., the value of notes issued at the home location falls as distance increases because the maturity of the debt increases. Condition (4.10) says that d_t^* is chosen to equate the marginal benefit of increased distance to the marginal cost of the capital loss associated with carrying the notes further away from home. The notes decline in value with distance leaving the buyer with less on hand to satisfy the cash-in-advance constraint, i.e., while consumption goods purchased further away "taste" better, a note carried further away drops in value as a function of d so fewer goods can be purchased. This is summarized in the following proposition.

PROPOSITION 3. At each date, t, there exists a critical distance, d_t^*, beyond which bank notes of banks at location d will not circulate.

The optimal distance depends on σ and leverage. Note prices which at various times are quoted in Philadelphia as "uncertain" (or which are note listed at all) may, at other times, be quoted because σ or bank leverage have changed.

19. If the volume of notes in transit were known, so that Geske's (1977) formula was appropriate, the same comparative statics would hold (Geske, 1977).

20. Note that if there are notes in transit then, in equilibrium, the outstanding amount of notes would be divided between notes in transit, notes sent for redemption, and notes held until next period.

For example, in Table 4.5, Arkansas and Nebraska are initially quoted, but subsequently are not quoted, even though the notes of more distant banks are quoted.

Now consider what happens if the household buyer goes to the home market and purchases goods from the household seller using bank notes from the home location, i.e. $d_t = 0$. Then, since the debt has no maturity, the option could be exercised instantly. If a bank note issued by a bank at the home location traded at discount at the home location, it could be costlessly converted into consumption goods at par as long as the bank is solvent. If the note were not priced at par, then this would occur until the bank was closed. Hence, the notes of banks at the home location must have no discount at the home location. By Proposition 2, if $d = 0$, then the discount is zero if the bank is solvent. Thus, $d = 0$ implies that those notes are risk free. Consequently, the notes of Philadelphia banks should always have a zero discount (which they do in the data).

During the Free Banking Era transportation costs and the duration of trips declined greatly with the spread of the railroad across the continent. This corresponds to an exogenous reduction in the time it takes to get back to a given location, i.e., to a reduction in d for a given location. Technical change reduces *d*, and hence increases notes prices (reduces discounts), *ceteris paribus.* But, if other factors change, while technical progress is occurring, then note discounts will not necessarily decline smoothly.

Note discounts are not monotonically increasing in time to return, d^*, because of the effects of risk (σ) and bank leverage. The factors which *a priori* evidence suggests affect bank risk are captured by σ. Coalitions of banks which may have effectively been self-regulating, in particular the Suffolk Bank system, encompassing the banks of New England, correspond, in the context of the above model, to a reduction of σ. Similarly, σ can be interpreted as capturing the effects of branching restrictions, insurance, and the default risk associated with bank issuance of additional money by wildcat banks, and whether or not the type of banking system, free or chartered matters.

A final feature of the equilibrium note prices is proven in Gorton (1996) in the context of the same model (but where σ is not exogenous). This feature concerns the modal discount. We state it here to explain the subsequent use of the modal discount in the empirical work. It would seem that the notes of different solvent banks at the same location could be priced differently at some particular distant location so long as the different prices reflected the different default risks. This would be true in efficient markets if notes were not used as a medium of exchange. The fact that notes are used as a medium of exchange, however, changes this intuition. Gorton (1996) shows this as given in the following proposition.

PROPOSITION 4. All solvent banks at the same location will have identical discounts at given distant locations and given date, t (assuming banks have the same leverage).

The proposition says that equilibrium requires all banks to choose their asset risk, σ, to be the same. While this is beyond the current model, the intuition for this result can be easily seen based on the above results. Consider the notes of two banks the same distance away, but with different risks (σ). A consumer holding notes of these two banks will not be indifferent between them even when their default risks are accurately priced. The reason is that if the consumer moves still further away from the issuing banks' location, increasing the time to redemption (maturity), the riskier banks' notes will decline in value by relatively more, hence purchasing less consumption units at the distant location. A less risky bank's notes will be preferred as a medium of exchange while the riskier bank's notes will be sent for redemption. But then equilibrium requires all banks with circulating notes to have the same risk and, hence, they are priced the same. This price is the modal discount.

4.5. THE BEHAVIOR OF BANK NOTE PRICES

If secondary note markets accurately priced risk, that is, accurately priced the redemption option, then the private money contract was enforceable in the sense that note holders would not suffer an unanticipated (i.e. unpriced) transfer to the note issuer (via issuance of additional currency as in wildcat banking or via increases in bank asset risk). The question to be addressed now is: Do bank note prices reflect bank risk?

To begin, a measure of bank risk is required. In the note pricing model, bank risk is completely captured by the variance or volatility of bank asset values. If bank notes can be priced with the Black-Scholes model as applied to corporate debt by Merton (1974), using the above result, then the volatilities of bank assets, i.e., σ's, implied by the note prices can be extracted be inverting the Black-Scholes formula. Using the closed-form Black-Scholes solution depends upon some strong assumptions. These are discussed below.

The volatility measure of risk is obtained from the note prices by inverting Eq. (4.10) for each state and date. Note that it is in this step that the importance of the redemption option and technological change enter the procedure. Leverage and trip time (i.e., maturity) are used in the formula to obtain the implied volatilities and do not enter the subsequent regressions. Technological change is captured in the calculation of the implied volatilities since maturity declines as transportation improves.

The method outlined above uses the exact closed form pricing solution for bank notes obtained in Proposition 2 under the assumption that there are no

notes in transit or that agents behaved as if there were no notes in transit.[21] Application of the Black-Scholes formula also requires assuming that the volatility and risk-free interest rate are constant through time. The first of these assumptions may be violated. Evidence suggests, however, that this violation is not likely to be important.[22] The second of these assumptions may also be violated. But, the implied returns on the bank notes are so high that the results are robust to a number of interest rate assumptions.[23]

The next step in empirically testing the model is to relate the measures of bank riskiness extracted from note prices to the measures of bank riskiness: the implied risk measures are regressed on the measures of bank riskiness discussed above. If secondary note markets functioned efficiently then the risk attributes of state banking systems should be priced. Explanatory variables, thus, include a dummy variable indicating whether the state is a member of the Suffolk System (*SUFFOLK*), a dummy variable indicating whether the state is a branch banking system (*BRANCH*), and a dummy variable indicating whether there is a state sponsored insurance arrangement (*INSURANCE*), and a dummy variable indicating whether the state is a free banking state (*FREE*).[24] (There are only a handful of risk variables available due to the data limitations associated with this period.) Finally, two variables capturing aggregate factors are included: a monthly index of stock prices (*SDEX*), and a dummy variable for the periods of suspension (*SUS*).

Table 4.7 reports the results of regressing the implied volatilities on the risk measures. Remarkably, the results in Table 4.7 are largely as expected. The R^2s are comparable to similar studies of modern bank debt (e.g. Flannery and Sorescu, 1995). The estimated coefficients on Suffolk system membership, branch banking, and insurance are all of the correct sign and significant. The presence of any of these factors is associated with lower volatility of bank assets (and hence lower discounts *ceteris paribus*). (This is true whether year dummies are included or not.)

21. The assumption that there are no notes in transit is made because there are insufficient data to make any other assumption.

22. The results of Schmalensee and Trippe (1978) and Latane and Rendleman (1976) demonstrate the value of using the Black-Scholes model to predict volatilities despite the inconsistency of using a model which assumes a constant variance to recover a possibly nonstationary variance. See Galai (1983) for further discussion.

23. A variety of interest rate assumptions were attempted. A series of annual commercial paper rates from Macaulay (1938) was used. Also, the risk free rate was, alternatively, exogenously set to zero and three percent for the period. No interest rate assumptions affects the results because the implied returns on the bank notes are so high.

24. The dummy variable is set to one when a state adopts free banking. In fact, such a state would have both free and chartered banks, but there is no feasible way to incorporate this information since it is not generally available.

Table 4-7. IMPLIED VOLATALITY REGRESSIONS ($N = 3384$)

Independent Variable	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**
Intercept	38.86[a]	37.79[a]	38.87[a]	37.82[a]	51.41[a]	49.10[a]	51.61[a]	49.31[a]
	(1.73)	(3.33)	(1.71)	(3.30)	(1.64)	(3.03)	(1.65)	(3.033)
Suffolk	−1.82[a]	−2.37[a]	−0.93	−1.49	−10.89[a]	−11.93[a]	−11.16[a]	−12.25[a]
member	(0.671)	(0.680)	(0.670)	(0.683)	(0.70)	(0.704)	(0.737)	(0.744)
Suspension	−11.32[a]	0.573	−11.33[a]	0.589	−14.56[a]	0.961	−14.50[a]	0.966
	(0.941)	(2.54)	(0.933)	(2.52)	(0.863)	(2.29)	(0.864)	(2.29)
Free banking	1.89[a]	0.77	—	—	−0.82	−2.58[a]	—	—
	(0.66)	(0.736)			(0.606)	(0.674)		
Good free	—	—	−0.43	−1.69	—	—	−0.54	−2.24[a]
			(0.717)	(0.791)			(0.656)	(0.720)
Bad free	—	—	8.37[a]	7.19	—	—	−1.78	−3.68[a]
			(1.05)	(1.09)			(1.04)	(1.07)
Branch	—	—	—	—	−17.11[a]	−17.50	−17.41[a]	−17.85[a]
banking					(0.76)	(0.752)	(0.805)	(0.797)
Insurance	—	—	—	—	−22.67[a]	−23.28[a]	−22.98[a]	−23.66[a]
					(1.10)	(1.10)	(1.13)	(1.13)
Stock index	−0.11[a]	−0.05	−0.12[a]	−0.05	−0.145[a]	−0.056	−0.145[a]	−0.056
	(0.19)	(0.05)	(0.019)	(0.051)	(0.017)	(0.046)	(0.017)	(0.046)
Year dummies	No	Yes	No	Yes	No	Yes	No	Yes
R^2	0.056	0.081	0.07	0.10	0.22	0.25	0.22	0.25
F-value	51.94	14.04	54.64	16.34	164.52	46.98	141.20	45.25
(Prob. > F)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)

NOTES: Standard errors in parentheses.
[a]Indicates significance at the 0.05 confidence level.

Free banking, however, does not appear to be associated with higher risk. In columns (1) and (2), where the branching and insurance factors are omitted, free banking increases risk, consistent with the traditional assertion, but is not significant in column (2). When the branching and insurance factors are included, columns (5) and (6), free banking lowers risk (but again is insignificant in one of the cases). Suppose a finer distinction is made, following Rockoff (1974b). On the basis of independent evidence Rockoff (1974b) suggests that the free banking states can be usefully divided into two groups: "good" free banking states and "bad" free banking states.[25] The results imposing this distinction, columns (3) and (4), and columns (7) and (8), still provide a mixed pattern of results. In column (8) both variables are significantly negative, but insignificantly different from one another.

That free banking was not perceived to be riskier is consistent with the evidence that wildcat banking was not common. The extensive commentary about wildcat banks by contemporaries of this period rarely distinguished between free banking states and chartered banking states. Moreover, many chartered banking systems were subject to abuse so that entry was not always difficult (see Chaddock, 1910; Hammond, 1957, pp. 332–37; Knox, 1903, p. 413; Ng, 1987; Sylla, 1985). It is also worth noting that, aside from New York, almost all of the entry into banking under free banking laws occurred in the 1850s, by which time the railroad and telegraph were widespread. One conjecture might be that by this point the redemption option was a powerful device for preventing risk-shifting.

Finally, notice that volatility rises when the stock market declines. The suspension variable is difficult to interpret since its sign depends on whether the year dummies are present or not. Though not reported, it is worth noting that seasonal dummies were always insignificant.

4.6. CONCLUDING REMARKS

Previous research indicates that wildcat banking was not a prevalent problem during the Free Banking Era. The reason for this may be that market participants could discipline banks by pricing factors that affected risk and via the contractual redemption option. Properly pricing risk means that a bank which set out to overissue notes would obtain a market price of zero on its notes. The contract device of the redemption option may have allowed note holders to run on banks which attempted to add risk. This paper has investigated whether note markets functioned in this way. Taking account of the redemption option, and the effects of technological change on this option, the above results are quite suggestive of

25. Following Rockoff the "bad" free banking states were identified as Michigan, Indiana, Illinois and New Jersey. The remaining free banking states were classified as "good."

the ability of market participants to price bank risk. The results also suggest that the type of banking system, free or chartered, was not the primary factor determining the relative risk of different banking systems. Other risk attributes appear to have been more important. This is consistent with previous findings.

4.7. FOR FURTHER READING

The following references are also of interest to the reader: Gorton, 1985; Rockoff, 1985; Rockoff, 1990; Rolnick, 1988.

REFERENCES

Black, F., Scholes, M., 1973. The pricing of options and corporate liabilities. Journal of Political Economy 81, 637–59.

Breckenridge, R.M., 1899. The Comtroller's objections to currency reform. Journal of Political Economy 7, 253–65.

Cagan, P., 1963. The first fifty years of the national banking system—An historical appraisal. In: Carson, D. (Ed.,), Banking and Monetary Studies, Richard D. Irwin Homewood, Illinois.

Callender, G., 1909. Selections From the Economic History of the United States. Ginn and Company, Boston.

Calomiris, C., 1989. Deposit insurance: Lessons from the record, economic perspectives. Federal Reserve Bank of Chicago Economic Perspectives (May/June), pp. 10–30.

Calomiris, C., Schweikart, L. 1988. Was the South Backward?: North-South Differences in Antebellum Banking During Normalcy and Crisis. Working paper.

Campbell, C., 1932. The development of banking in Tennessee, Ph.D. Thesis, Vanderbilt University.

Carothers, N., 1930. Fractional Money: A History of the Small Coins and Fractional Paper Currency of the United States. Augustus M. Kelley, New York; 1967 reprint of original.

Chaddock, R., 1910. The Safety-Fund Banking System in New York State, 1829–1866. Government Printing Office, Washington, DC.

Cleaveland, J., 1857. The State Banking System of the State of New York, 1829–1866. Government Printing Office, Washington DC.

Dewey, D., 1910. State Banking Before the Civil War. Government Printing Office, Washington, DC.

Dillistin, W.H., 1949. Bank Note Reporters and Counterfeit Detectors, 1820–1866. American Numismatic Society, New York.

Duboff, R., 1980. Business demand and the development of the telegraph in the United States, 1844–1860. Business History Review 54 (4), 459–79.

Duboff, R., 1983. The telegraph and the structure of markets in the United States, 1845–1890. Research in Economic History 8, 253–77.

Duboff, R., 1984. The telegraph in nineteenth century America: technology and monopoly. Comparative Studies in Society and History 26 (4), 571–86.

Fishlow, A., 1964. Antebellum interregional trade reconsidered. American Economic Review 54, 352–64.
Fishlow, A., 1965. American Railroads and the Transformation of the Ante-Bellum Economy. Harvard University Press, Cambridge, MA.
Flannery, M., Sorescu, S., 1995. Evidence of bank market discipline in subordinated debenture yields: 1983–1991. University of Florida. Working paper.
Fogel, R., 1964. Railroads and American Economic Growth. Johns Hopkins, Baltimore, Maryland.
Friedmon, M., 1959. A Program for Monetary Stability. Fordham University Press, New York.
Friedmon, M., 1986. The Cost of Irredeemable Paper Money, Journal of Political Economy (June).
Friedmon, M., Schwartz, A.J., 1986. Has Government Any Role in Money?. Journal of Monetary Economics 17, 37–62.
Friedman, M., 1987. Monetary policy: Tactics versus strategy. In: James, A.D., Schwartz, A.J. (Eds.), The Search for Stable Money. University of Chicago Press, Chicago.
Galai, D., 1983. A survey of empirical tests of option-pricing models. In: Brenner, M. (Ed.), Option Pricing, Lexington, Mass.
Geske, R., 1977. The valuation of corporate liabilities as compound options. Journal of Financial and Quantitative Analysis 12, 541–52.
Gorton, G., 1985. Clearinghouses and the origin of central banking in the US. Journal of Economic History 45, 277–83.
Gorton, G., 1989a. Self-Regulating Bank Coalitions, The Wharton School, University of Pennsylvania, working paper.
Gorton, G., 1989b. An Introduction to Van Court's Bank Note Reporter and Counterfeit Detector. The Wharton School, University of Pennsylvania. Working paper.
Gorton, G., 1989c. Ante Bellum Transportation Indices. The Wharton School, University of Pennsylvania. Working paper.
Gorton, G., 1996. Reputation formation in early bank note markets. Journal of Political Economy 104, 346–97.
Grant, J., 1857. A Treatise on the Law Relating to Banking. T. & J.W. Johnson & Co, Philadelphia, Pennsylvania.
Hammond, B., 1957. Banks and Politics in America. Princeton University Press, Princeton, New Jersey.
Harding, W., 1895. The State Bank of Indiana. Journal of Political Economy 3, 1–36.
Latane, H.A., Rendleman, R.J., 1976. Standard deviations of stock price ratios implied by option prices. Journal of Finance 31, 369–81.
Lindstrom, D., 1975. Demand, markets, and eastern economic development: Philadelphia, 1815–1840. Journal of Economic History 35, 271–73.
Lucas, R., 1980. Equilibrium in a pure currency economy. In: Kareken J.H., Wallace, N. (Eds.), Models of Monetary Economics, Federal Reserve Bank of Minneapolis, pp. 131–46.
Macaulay, F., 1938. The Movements of Interest Rates, Bond Yields, and Stock Prices in the United States Since 1856. National Bureau of Economic Research, New York.
Mercer, L., 1982. The antebellum interregional trade hypothesis: a reexamination of theory and evidence. In: Ransom, R. Sutch, R., Walton, G. (Eds.), Explorations in the New Economic History. Academic Press.

Merton, R., 1974. On the pricing of corporate debt: The risk structure of interest rates. Journal of Finance 29, 449–70.
Mullineaux, D., 1987. Competitive monies and Suffolk bank system: A contractual perspective. Southern Economic Journal 53, 884–99.
Ng, K., 1987. Free Banking Laws and Barriers to Entry in the Banking Industry. Working paper, California State University at Northridge.
Pred, A., 1980. Urban Growth and City-Systems in the United States, 1840–1869. Harvard University Press, Cambridge.
Rockoff, H., 1974a. Money prices and banks in the Jacksonian era. In: Fogel, R., Engerman, S. (Eds.), The Reinterpretation of American Economic History. Harper and Row, New York.
Rockoff, H., 1974b. The free banking era: a reexamination. Journal of Money, Credit and Banking 6, 141–67.
Rockoff, H., 1975. The Free Banking Era: A Reconsideration. Arno Press, New York.
Rockoff, H., 1985. New evidence on free banking in the United States. American Economic Review 76, 886–89.
Rockoff, H., 1989. Lessons from the American experience with free banking. National Bureau of Economic Research Working Paper on Historical Factors in Long-run Growth, No. 9.
Rockoff, H., 1990. The capital market in the 1850s. National Bureau of Economic Research Working Paper on Historical Factors in Long-run Growth, No. 11.
Rolnick, A., Weber, W., 1982. Free banking, wildcat banking and shinplasters. Quarterly Review, Federal Reserve Bank of Minneapolis (Fall).
Rolnick, A., 1983. New evidence on the free banking era. American Economic Review 73, 1080–1091.
Rolnick, A., 1984. The causes of free bank failures. Journal of Monetary Economics 14, 267–91.
Rolnick, A., 1988. Explaining the demand for free bank notes. Journal of Monetary Economics 21, 47–72.
Rubinstein, M., 1976. The valuation of uncertain income streams and the pricing of options. Bell Journal of Economics 7, 407–25.
Schmalensee, R., Trippi, R.R., 1978. Common stock volatility expectations implied by option premia. Journal of Finance 32, 129–147.
Schmidt, L., 1939. Internal commerce and the development of a national economy before 1860. Journal of Political Economy 47, 798–822.
Schweikart, L., 1987. Banking in the American South from the Age of Jackson to Reconstruction. Louisiana State University Press, Baton Rouge, LA.
Sumner, W.G., 1896. A History of Banking in the United States. New York.
Svensson, L.E.O., 1985. Money and asset prices in a cash-in-advance economy. Journal of Political Economy 93, 919–44.
Sylla, R., 1985. Early American banking: the significance of the corporate form. Business and Economic History 14, 105–23.
Thompson, R.L., 1947. Wiring a Continent: the History of the Telegraph Industry in the United States, 1832–1866. Princeton University Press, Princeton.
Van Vleck, G., 1943. The Panic of 1857. Columbia University Press, New York.
Whitney, D.R., 1878. The Suffolk Bank. Riverside Press, Cambridge rated.

5

The Development of Opacity in U.S. Banking*

GARY B. GORTON

5.1. INTRODUCTION

The financial crisis of 2007–2008 has led to widespread calls for "transparency." Some blame the crisis on a lack of transparency. But, banking is inherently opaque. Were it not opaque it could not be able to produce money. Bank opacity requires that banks be regulated and examined. Calling for "transparency" is misguided.

To those calling for "transparency" the financial system and some financial instruments are too opaque and too complex. This may explain why regulators were unaware of the extent of the shadow banking system and of the widespread use of some newer financial instruments. This lack of awareness, in the eyes of these observers, can be remedied by more transparency. The reasoning for this is the familiar—if vague—notion that "sunshine is the best disinfectant." In fact, the Dodd-Frank Act was in large part motivated by this concern; it mandated many new types of financial disclosure and created an independent office, the Office of Financial Research, which is charged with collecting data and empowered to obtain new data in order to inform the newly created Financial Stability Oversight Council.

There can be some appeal to economics as well, in a very general sense. Economics implicitly suggests that transparency is good. For example, the idea that opening financial markets is good and closing them is bad comes from the theory

* This paper draws on ideas in joint work with Tri Vi Dang, Bengt Holmström and Guillermo Ordonez to whom I am very grateful. Thanks to Yiming Ma and Arwin Zeissler for research assistance.

of complete markets. It is best when markets are complete, that is, when the existing financial claims are sufficient to construct any conceivable payoff, corresponding to any state of the world (in a world without frictions). Complete markets are desirable because risks can be hedged. A separate idea about financial securities is that "market efficiency" is desirable. This says that the prices of financial securities are informative, providing information to investors, for example.

The existence of informative financial markets allows for markets can discipline banks, but that requires information and that's a problem. First, with deposit insurance depositors have no incentive to differentiate good banks from bad banks. Even for stock investors, there is little information. Bartlett (2012) writes: "The problem with prevailing bank disclosures . . . is that they are generally limited to aggregated metrics that make it difficult to assess a bank's credit concentrations, underwriting standards, or portfolio quality . . . The second factor relates to the complexity of a bank's investment activities" (pp. 298–99).

It is more than a lack of transparency, secrecy surrounds banks. Most of the infrastructure surrounding banks is precisely intended to make them opaque to outsiders. The results of bank examinations are kept confidential by the regulators and borrowing from the discount window is (supposed to be) kept secret. Even important special information is not revealed. John Carney of CNBC: "The Senate report on JP Morgan Chase's London Whale fiasco revealed that federal regulators secretly downgraded the bank's management rating last summer—a fact kept from investors and the public until last week" (NetNet, CNBC, March 20, 2013). And the financial instruments banks created, like subprime mortgages, are complex and opaque. During the crisis the Federal Reserve System did not reveal which institutions got emergency loans. And so on.

Banks are special. They face runs and that is why they are regulated. At the root of this specialness is the fact that banks are optimally opaque. So, contrary to the pleas for transparency, in this paper I argue that banks are opaque for a good reason and this is why they are regulated. I look at U.S. financial history and show that the production of private money by banks optimally involves closing informative financial markets where bank liabilities (debt and equity) are traded. The efficient use of bank claims as money entails eliminating informative financial markets, so that banks are opaque and their monies consequently are accepted at par. My argument is that banks are supposed to be opaque. But, that makes them vulnerable to runs and hence they are regulated. A call for transparent banks is oxymoronic. One must start with the question of why banks exist, what is that they do, and why are they so different that they need to be regulated.

The output of a bank is its debt which is used as money, whether demand deposits, private bank notes, sale and repurchase agreements, or other forms of short-term debt. A "bank" is a firm which issues short-term debt in whatever form. For short-term bank debt to function efficiently as money it must trade at

par, keep its value, that is, it must be accepted at face value without any suspicion that it is worth less than its face value. And no information should become available to create suspicion. For this to be successful banks keep secret the value of the backing for their debt. Banks, for example, lend predominantly to households and small businesses, entities for which there is little or no public information. Bank examiners check the banks' portfolios, but their assessments are also kept secret.

This opacity has a cost: short-term bank debt is vulnerable to bank runs because the backing for bank debt is not riskless. The private sector cannot create riskless assets. In a bank run, the holders of the debt become suspicious about the backing of the debt. A financial crisis is an information event, occurring when holders of bank debt become so suspicious of the backing of the debt that they seek to obtain their cash back en masse. Obviously the banking system cannot honor these demands and so the banking system is insolvent. This occurs when there is unexpected news of a coming recession or unexpected news of a decline in an important sector of the economy. Hence the conundrum: the business of banking inherently requires opacity, but that can create runs. This is why banks are regulated and examined.

In order to understand the above points, I focus on an example. I trace the historical transition from private bank notes to demand deposits in the United States. Rather than make the above points in theory, U.S. financial history is used to show how this endogenously occurred and made the economy more economically efficient.[1] The transition I focus on is one example, but suffices to make the point.

Before the U.S. Civil War, the predominant form of bank liabilities used as money was private bank notes. The federal government did not issue paper currency at that time, but banks issued their own paper currencies. Bank notes traded at discounts from face value, revealing information about the issuing banks' backing assets. And, bank equity traded in information-revealing stock markets. Gradually, demand deposits (checking) grew significantly and after the Civil War the U.S. government imposed a tax on private bank notes, essentially forcing them out of existence. The transition from bank notes to demand deposits is instructive about the optimal form of banking and bank money. The transition involved closing informative bank note and stock markets in which bank liabilities traded, reducing the available

1. The corresponding theory can be found in Gorton and Pennacchi (1993), Holmström (2008, 2011), Dang, Gorton, and Holmström (2012) and Dang, Gorton, Holmström, and Ordonez (2013). These papers make the case that the optimal transaction medium is debt because debt minimizes the incentive to produce private information which can lead to adverse selection when the private money is used to trade. In order to privately create such money, banks are opaque. While this is socially optimal, it can lead to runs, which is why banks are regulated.

information, so that demand deposits could more effectively function as money. The transition involved the creation of opaque banks, not via regulation but endogenously.

Closing private bank note markets and bank stock markets was possible because a monitoring role developed centering on private bank clearing houses. Ostensibly founded to clear checks, internalizing the bank note secondary market, clearing houses produced information about member bank risk, without revealing (most of) it. During financial crises—bank runs—clearing houses assumed the role of a central bank, acting as a lender-of-last-resort. During a crisis, the clearing house managed the information environment, further suppressing information about member banks while at the same time producing information that it kept secret when the clearing house examined some banks during a crisis. The clearing house also issued new liabilities, which were the joint liabilities of the member banks. These two acts, suppressing bank-specific information and issuing joint liabilities, effectively joined the members into a single banking system. Rather than focusing on whether any specific bank was weak, the clearing house by these two acts, made the only relevant question one of whether the banking system was solvent.

The idea that firms or other nonmarket organizations may be better than markets in allocating resources is hardly a new idea (see, e.g., Coase (1937), Williamson (1975), and Holmström (1999)). What is different about banks is that the attendant financial markets must be shut down to produce efficient private money. And this causes private bank clearing houses to assume the role of suppressing information, but also to assume a central bank-like role during financial crises and in non-crisis times. The clearing house is a unique organization—not a firm—necessary because bank-specific information had to be suppressed in order for banks to produce money. The origin of the Federal Reserve System lies in these private bank clearing houses, in large part.

In the context of the above ideas, the financial crisis of 2007–2008 is also briefly discussed. I discuss what happened during the crisis and then I focus on three particular informational aspects of the crisis. This is followed by the conclusion.

5.2. PRIVATE BANK NOTES

It is perhaps easiest to understand the above information issues with bank money by starting with the period of U.S. history when banks issued their own currency, 1837–1863, sometimes referred to as the Free Banking Era. This was a period, prior to the U.S. Civil War, during which the U.S. government did not issue paper money. It was also a period in which the use of demand deposits (checking accounts) was growing. I focus on the transition from private bank

notes to demand deposits, and the concomitant alterations in the information environment concerning banks.[2]

A private bank note was a perpetual noninterest-bearing liability of a specific private bank. The note holder had the right to go back to the issuing bank at any time and demand redemption in gold or silver. The notes were printed in denominations similar to government money today, e.g., one dollar bills, five dollar bills, etc. During 1837–1863 there were around 1,500 different banks' currencies circulating at one time. Since these were the liabilities of private banks, these currencies were not riskless, so when they circulated at any distance from the issuing bank—so that returning to redeem the money would take time—the notes circulated at discounts. For example, the bank notes of Boston banks would circulate at discounts from par in New York City. A ten dollar note of a particular bank in Boston would circulate in New York City at say a five percent discount from face value; a ten dollar note might only buy $9.50 worth of goods in New York City.

Bank notes of nearby banks, say the notes of Boston banks in Boston, would have no discount. A note holder of a Boston bank could always go back to the bank and ask for gold, without bearing any real transportation costs and without taking much time; the bank was viewed as riskless over very short intervals of time. But, outside Boston there would be discounts on the notes' face values, and the discounts increased as the distance from the issuing Boston bank increased. Over time, discounts decreased as technological change occurred, i.e., the introduction of the railroad, which made it easier to return to the Boston bank, for example. At a distance away from the issuing bank, a transaction would be made at the note discount. The discount was determined in informal note secondary markets in which note brokers traded bank notes. The discounts were recorded by newspapers called "bank note reporters," the financial press of the time. (See Dillistin (1949).) A Philadelphia bank note reporter, for example, *Van Court's Counterfeit Detector and Bank Note List*, covered 3,089 banks in 35 states, territories, and provinces of Canada. See Gorton (1989).

So, in order to transact with a customer, a storekeeper would look up the discount in the local bank note reporter. The banknote reporter, usually published monthly, got the discount information from a note broker (who traded in an informal note market). Each large city had at least one bank note reporter. The bank note reporter would list the discounts on all bank notes circulating in that particular location, say in Boston or New Haven. Notes from very distant locations would not circulate, e.g., notes of Wyoming banks did not circulate in

2. Not all states passed Free Banking laws, though banks in all states issued private currency. For background on the U.S. Free Banking Era see Rockoff (1975), Rolnick and Weber (1983, 1984), and Gorton (1996, 1999).

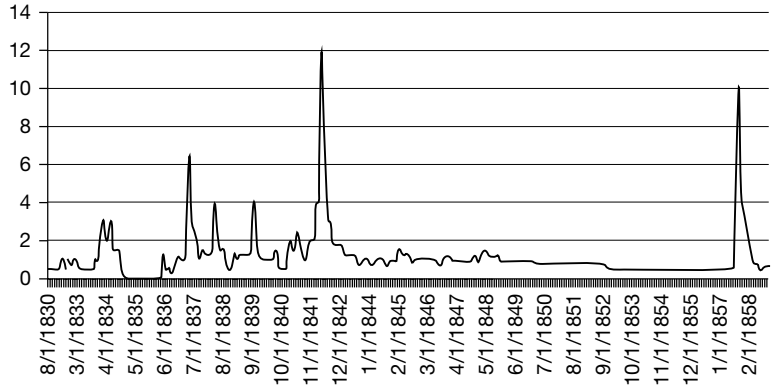


Figure 5.1 Bank of Virginia Note Discounts in Philadelphia (% from face value)
SOURCE: Gary B. Gorton and Warren Weber, "Quoted Discounts on State Bank Notes in Philadelphia, 1832–1858," Research Department, Federal Reserve Bank of Minneapolis. http://minneapolisfed.org/research/economists/wewproj.html.

Philadelphia. Most notes from Boston banks circulated at the same discount in Philadelphia, but not all Boston banks. And the further away the location of the banks, the less like that the notes in a distant location would circulate at the same discount.

Figure 5.1 shows the discounts in Philadelphia on a single bank, the Bank of Virginia, over time. The *y*-axis scale is percent discount. Most of the time the discount is low, but there is a lot of volatility to the discount. In the face of such volatility, the bank note reporter's role was to provide market participants with the discount at the time of the transaction. Table 5.1 further provides a sense of the variation in note discounts. As examples, the table shows the average annual discount, in Philadelphia, averaged over months, together with the number of banks for Ohio and for South Carolina. (See Gorton (1989).) The mean discounts and their standard deviation in Ohio are both much larger than those of South Carolina. State banking systems were regulated differently, so the risk could differ even holding distance constant. It is clear that the discounts varied over time, sometimes rather dramatically. The standard deviation also varies over time.

The bank note market was efficient, in the financial economics sense of "market efficiency," i.e., security prices contain and reveal information. Here it means that the discounts on notes some distance from the issuing bank accurately reflected the bank's risk, given that it would take time to get to that bank (the effective maturity of the note), time during which the bank could fail. See Gorton (1999). Furthermore, the discounts functioned to discipline new banks. The discounts of new banks were higher than the discounts on the notes of other banks at the same location, creating an incentive for note holders to go back and demand cash, to check on the new bank. The higher

Table 5-1. NOTE DISCOUNTS ON OHIO AND SOUTH CAROLINA NOTES IN PHILADELPHIA

Year	Ohio			South Carolina		
	Mean Discount	Standard Deviation	Number of Banks	Mean Discount	Standard Deviation	Number of Banks
1839	4.18	1.33	38	3.57	1.65	11
1840	4.76	1.55	42	0.34	0.83	12
1841	7.45	3.44	40	1.27	0.91	12
1842	14.18	13.32	34	2.54	1.49	12
1843	14.4	20.18	36	1.81	0.59	12
1844	10.49	16.96	35	0.94	0.25	12
1845	8.97	14.24	35	1.26	0.21	12
1846	7.68	13.97	40	1.35	0.33	13
1847	8.26	18.23	39	1.00	0.37	13
1848	9.18	19.01	44	1.78	0.96	15
1849	12.16	23.23	44	1.17	0.63	15
1850	12.84	24.17	44	0.85	0.26	14
1851	12.4	23.96	43	0.84	0.33	14
1852	6.16	17.91	30	0.87	0.26	14
1853	2.63	10.27	39	0.75	0.11	16
1854	1.86	0.86	37	0.96	0.19	17
1855	3.08	8.18	37	1.08	0.35	18
1856	2.64	8.21	38	0.83	0.11	18
1857	5.69	12.12	38	1.97	2.65	19
1858	6.5	16.33	36	1.63	1.12	20

SOURCE: *Van Court's Counterfeit Detector and Bank Note List* (see Gorton (1989)).

discount thus acted to reward those monitoring new banks. See Gorton (1996). In sum, bank note markets functioned as "efficient" markets; the discounts were informative about bank risk. Banks at the same location competed, and the note market enforced common fundamental risk at these banks.

While the note market was efficient from the point of view of the note discounts, there was a market failure: it was not economically efficient (i.e., the best allocation of goods and services could not be made based on transaction with these notes). The problem was that the costs of transacting with bank notes were high. Sumner (1896) explains this in his *History of Banking*:

> The bank-note detector did not become divested of its useful but contemptible function until the national bank system was founded [creating government money]. It is difficult for the modern student to realize that there were hundreds of banks whose notes circulated in any given community. The bank-notes were bits of paper recognizable as a species by shape, color, size and engraved work. Any piece of paper which had these came with the prestige of money; the only thing in the shape of money to which the people were accustomed. The person to whom one of them was offered, if unskilled in trade and banking, had little choice but to take it. A merchant turned to his 'detector.' He scrutinized the worn and dirty scrap for two or three minutes, regarding it was more probably 'good' if it were worn and dirty than if it was clean, because those features were proof of long and successful circulation. He turned it up to the light and looked through it, because it was the custom of the banks to file the notes on slender pins which made holes through them. If there were many such holes the note had been often in bank and its genuineness was ratified. All the delay and trouble of these operations were so much deduction from the character of the notes as current cash. A community forced to do its business in that way had no money. It was deprived of the advantages of money. We would expect that a free, self-governing, and, at times, obstreperous, people would have refused and rejected these notes with scorn, and would have made their circulation impossible, but the American people did not. They treated the system with toleration and respect. A parallel to the state of things which existed, even in New England, will be sought in vain in the history of currency. (p. 455)

These complaints were commonplace during the Free Banking Era.

Thus, although the discounts displayed individual bank risk, there was a market failure in terms of private banks being able to produce debt that could be used as money without the concomitant disadvantages of bank notes. Bank notes were not an efficient transaction medium.

5.3. DEMAND DEPOSITS AND BANK STOCKS

Demand deposits (checking) were a financial innovation that grew enormously during the years before the U.S. Civil War; see figure 5.2. Checking accounts had several advantages over private bank notes. First, these accounts paid interest. And, second, there was no discount on local checks; the checks were accepted at the value the payer denominated. The disadvantage is that checks not only depend on the bank but also on the person writing the check, who must have the money in the bank account. A check is a "double claim," being a claim on both a specific bank and a specific person's account. Consequently, markets for such specific claims would be very thin; it would be too costly to have a secondary market in the checks of individual people at their specific banks. So checks first grew in urban areas where a person's identity was most easily verified. One way to think of the discount on checks is that the discount was either zero or 100 percent. Out-of-town checks had a 100 percent discount at first, while local checks had zero discounts. It took some time for out-of-town checks to become accepted.

Bank note markets were organized informally by note brokers. But, checks require "clearing." The checks written on one bank would be deposited at another bank. So, the receiving bank had to present the check to the other bank for payment. With many checks, the process of clearing by banks each sending messengers to all the other banks to present checks for payment, while all the

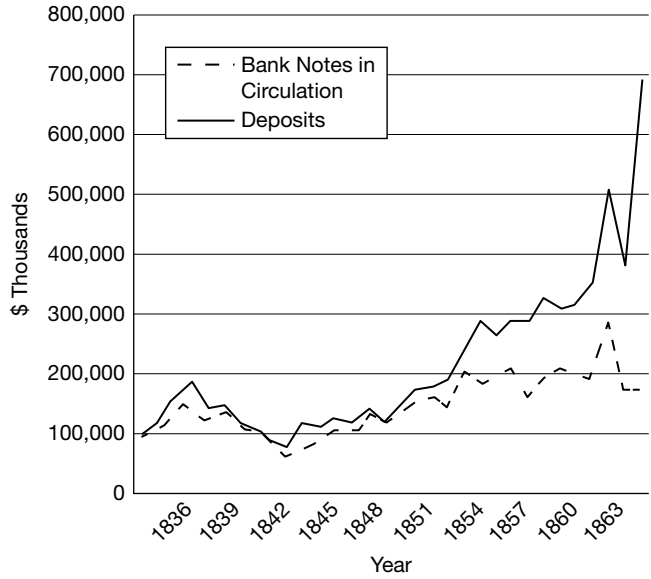


Figure 5.2 Growth of Demand Deposits
SOURCE: *Historical Statistics of the United States 1789–1945* (1949), pp. 263–4.

other banks were sending their messengers with checks for payment, was quickly very cumbersome. Clearing houses were established so that banks could go to a single location and submit and receive checks from other banks. The first clearing house in the United States was established in New York City in 1853 and subsequently spread across the nation (see Gibbons (1859), Cannon (1910), Redlich (1951), Gorton (1984, 1985a), Timberlake (1985), and Gorton and Mullineaux (1987)).

In the clearing process a bank might face another bank which owed the first bank a large amount of money or vice versa on another day. This counterparty risk, as we would call it today, meant that the clearing house took on other responsibilities related to monitoring member banks. Clearing houses imposed capital requirements, reserve requirements, interest rate restrictions, ongoing audits and reporting. (See Cannon (1910).) In the process of clearing, the clearing house became informed about the state of individual member banks and, in fact, started a bank examination process. The results of examinations were kept secret, but the clearing house did require members to publish balance sheet information weekly in newspapers.[3] As Bolles (1903, p. 379) explained:

> The extent of the supervision exercised by this association over its members the public will never know, because it is best that much of it remain secret. The banks thus associated learn more about one another than they ever would if acting entirely alone and examinations are made, and warnings given, of which the public has no knowledge. The direct interest that every bank has in knowing the true condition of every other member is one of the great merits of the system.

The clearing process produced information, as did clearing house member bank examinations, but other than the information that was required to be made public, no other information was revealed. In other words, because there were no discounts to the face value of demand deposits, and because the information garnered by the clearing house was not made public, information from note discounts was effectively lost to the public. But it was still produced and the clearing house acted on this information.

In order for checks to be accepted at par, that is with no discount for the risk of the issuing bank, there must be no information available to price the bank risk of a bank's checks. If the two parties to the transaction understand that neither party has any secret information about the risk of the bank such that the uninformed party is taken advantage of, then the check will trade at par. I will accept your check for $100.00 in exchange for $100.00 of goods.

3. On clearing house bank examinations see Bolles (1903), Cannon (1910), and Smith (1908). Smith (1908) described the government bank examinations as "defective."

Clearing houses replaced bank note markets and kept the information about the risk of individual banks secret. But, what about bank stock prices, renowned as information-revealing prices in an efficient market? Bank stock prices, which in the Free Banking Era were publicly available in New York City for large banks, would reveal information, because the stock prices were efficient. Such information-revealing prices would reveal information about bank risk and could have led to discounts on checks or runs on banks. Why did that not occur?

The answer is quite straightforward: the market for bank stocks was also effectively closed, by the banks themselves. Banks took actions to make their stocks very illiquid. Goetzmann, Ibbotson and Peng (2001) collected individual firm stock prices for NYSE stocks over the period 1815–1925. They exhaustively collected stock prices from a variety of sources, covering over 600 companies during the sample period. Their data display an interesting phenomenon, which is portrayed in figure 5.3. The figure graphs the total number of companies with actively traded stock in their sample, and the total number of banks in the sample with traded stock. Bank stocks were quite prevalent up to 1872 after which they disappear.

Banks remained public companies but they took actions to insure that their stock was illiquid. This was accomplished by making the stock price of a single share very high, out of reach of most investors. And, the stock ownership was

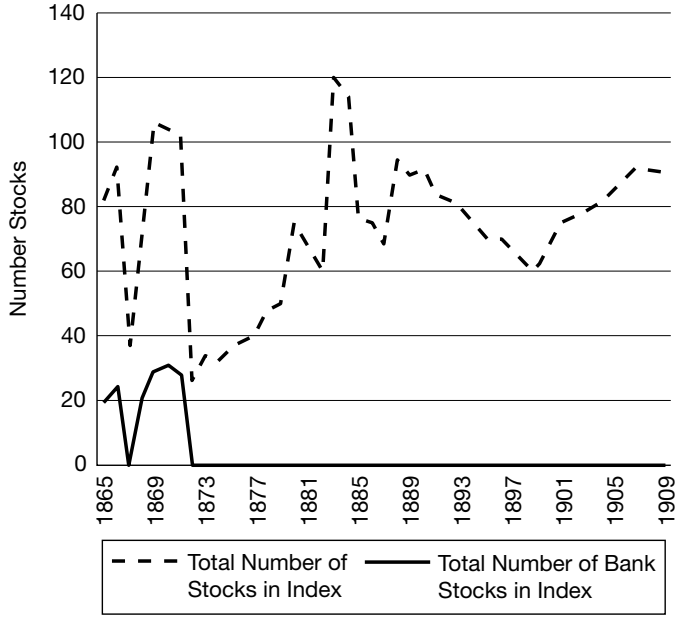


Figure 5.3 New York Stock Market, 1863–1909
SOURCE: Goetzmann, Ibbotson and Peng (2001).

concentrated. Loeser (1940, p. 158): "For a long time the stocks of these institutions [bank, trust, and insurance companies], particularly the leading ones, were looked upon as 'rich men's investments.' . . . In other instances there was a high degree of concentration of shareholdings among family groups and groups of business associates and other with allied interests."

Banks recognized that fluctuations in stock prices, in particular declines in a stock price, could lead to bank runs because the informative price could reveal that the backing assets had declined in value. As Stevenson (1910) put it:

> No bank can long exist without a complete trust on the part of the depositors. If stories which affect the bank's standing and character seem to be a part of the speculative tactics, should they grow, which may cause panic, then it is incumbent that those in the management of large banks see to it, as far as in their power to, and prevent the dealing of bank stocks and their quotations on the stock exchanges of the country. (p. 341)

Also Loeser (1940) noted:

> Within the past decade, with one exception, leading banks with issues listed in New York had their issues removed from listing. Many banks in other cities also delisted their securities. The reason generally given for this voluntary delisting was that the banks were apprehensive that the publicity which might be given to prices declines of their issues on the exchanges might be misconstrued by the public and might affect the confidence of depositors adversely (pp. 160–161).

The transition to demand deposits entailed making bank stocks illiquid, so that their prices would be uninformative. O'Sullivan (2007): "For the most part, bank stocks were not widely traded" (p. 517).

The Federal Reserve System was founded in 1914 with the express purpose of preventing banking panics. Indeed, it did prevent a panic in 1920 (see Gorton (1988) and Gorton and Metrick (2013)). For a brief period in the 1920s some banks listed on the New York Stock Exchange, as follows:

Bank of America, 1927–1928
Bank Manhattan, 1927–1928
Bank of New York, 1927–1929
Chase National Bank, 1927–1928
Chatham Phoenix National Bank, 1927–1928
Chemical National Bank, 1927–1928
Commerce Guardian Trust & Savings Bank, 1927–1929

Continental Bank, 1927–193
Corn Exchange National Bank, 1927–195
Farmers Loan & Trust, 1927–1928
Hanover National Bank, 1927–192
National City, 1927–1928
National Park, 1927–1929

But the banks quickly delisted in a few years. The Corn Exchange is the only bank that remained listed after January 1930.

The lack of information about banks persisted, even after deposit insurance was adopted in 1934. In 1964 the U.S. House of Representatives commissioned a study on the issue of bank opacity as it related to bank equity holders. The committee noted that:

> Stockholders of banks in many cases receive little or no information concerning the financial results of their bank's operations. Less than 50 percent of all banks publish annual reports. Of those who publish annual reports, 29 percent do not reveal the size of their valuation reserves. Before-tax earnings are not disclosed by 36 percent of all banks and after tax earnings are not disclosed by 34 percent of all banks.
>
> (U.S. House of Representatives (1964), p. v)

The report contained table 5.2 below. The table shows the number of shares traded in 1962 for different number of shares outstanding. Surprisingly, the number of shares traded monotonically declines in number of shares outstanding. In other words, larger banks with more shares outstanding have the lowest number of shares traded. The total annual trading volume of bank shares on the New York Stock Exchange is shown in figure 5.4. Until the early 1960s bank stock did not actively trade.

In the transition from bank notes to demand deposits two information-revealing markets closed: the market for bank notes which set the discounts; and bank stock markets. Closing information-revealing markets that would reveal bank risk was economically efficient because bank liabilities could then be accepted at par, avoiding the transactions costs associated with bank notes. However, this does not mean that information should not be produced, to distinguish good banks from bad banks. It means that is it the job of the bank regulators to do this.

Demand deposits were the "shadow banking" system of the National Banking Era, 1863–1914. It was thought that panics would end once the government entered the business of paper currency during the Civil War. But, panics continued with runs when people suspected the backing of the checking accounts. Economists and regulators were not sure of the extent to which checks were used as a transaction medium, and panics persisted until deposit insurance.

Table 5-2. NUMBER OF SHARES TRADED IN 1962 VERSUS TOTAL NUMBER OF SHARES OUTSTANDING AT YEAR END 1962

Shares Traded	**Number of Outstanding Shares**							**Total**
	0 to 10,000	**10,000 to 50,000**	**50,000 to 100,000**	**100,000 to 500,000**	**500,000 to 1,000,000**	**1,000,000 to 5,000,000**	**Over 5,000,00 0**	
Less than 1,000	51,684	15,816	1,372	384	98	120	40	69,514
1,001 to 50,000	719	4,288	2,288	2,837	217	30	–	10,379
50,001 to 100,000	–	–	28	177	242	19	–	466
100,001 to 500,000	10	–	–	74	166	299	–	549
500,001 to 1,000,000	–	–	–	–	–	60	30	90
More than 1,000,000	–	–	–	–	–	–	60	60
Total	52,413	20,104	3,668	3,472	723	600	130	81,110

SOURCE: U.S. House of Representatives (1964).

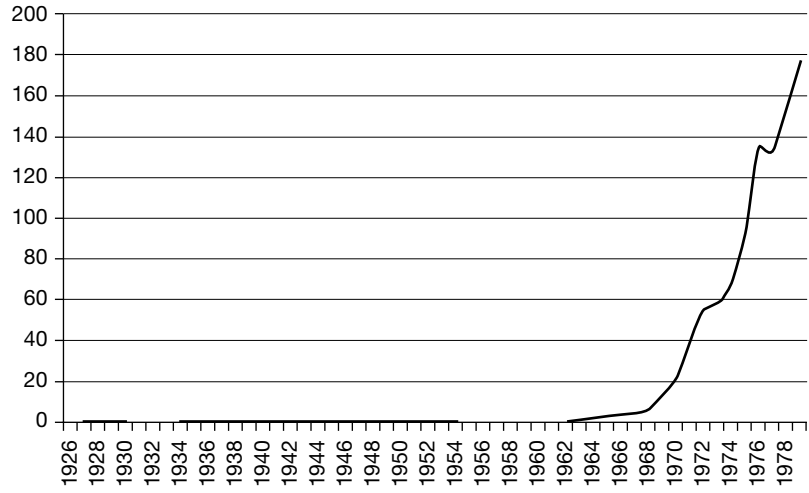


Figure 5.4 Bank Total Annual Trading Volume (NYSE Only, Millions of Shares, 1926 to 1979)
SOURCE: CRSP; SIC = 6010, 602x; EXCHCD = 1; SHRCD = 10, 11.

5.4. CLEARING HOUSES AND PRE-FED FINANCIAL CRISES

Before the Federal Reserve came into existence, financial crises were managed by the private bank clearing houses, which acted as lenders-of-last-resort. A financial crisis is a bank run; holders of bank short-term debt no longer want to hold the debt and instead want their cash back. The debt holders want cash because they have received information about a coming recession during which their bank may fail (see Gorton (1988)). Since banks are opaque there was no way for them to know which banks were weak and which were not. (See Gorton (1985b).) Hence, all banks were run on. When this happens in the entire banking system at the same time, banks cannot possibly honor the demands for cash because their assets—mostly loans—cannot be sold. The entire banking system is insolvent because the debt holders' contractual right to ask for cash cannot be honored.

The first act of the clearing house when a crisis started was to cut off the publication of bank-specific information, which was required in normal times, usually followed by suspension of convertibility, that is banks would refuse to pay cash to redeem checks.[4] During normal times, the clearing house required members to publish balance sheet information; newspapers published these numbers weekly. Bank-specific information might identify the weaker banks, which would then be subject to runs. To stop the desire to run on the banks, the clearing house had to convince bank debt holders that

4. This was always illegal but never enforced; see Gorton (2012).

the member banks were solvent, that the bank assets were illiquid but not in default. This required management of the information environment in two very specific ways. First, a securities market had to be *created* to reveal information about the solvency of all member banks jointly, effectively the banking system.[5] Secondly, the clearing house needed to convince the public that certain specific banks, those subject to persisting rumors of weakness, were in fact, solvent.

The clearing house also had to address the illiquidity problem. After suspension occurred, the clearinghouse issued "clearing house loan certificates," a new form of private money that could be used in the clearing process instead of cash.[6] Loan certificates were the jointly liability of clearing house members. In other words, the banks banded together formally by assuming this joint liability. The prospect of this happening meant that in normal times the member banks had an incentive to monitor each other. (See Gorton and Huang (2006).)

Individual member banks would apply to a clearing house committee for loan certificates, offering collateral from their balance sheets. The clearing house went to great lengths to protect the secrecy of which banks borrowed loan certificates. Preventing leaks concerning the loan certificate borrowings of individual clearing house members was important for preventing signs of weakness at banks with large borrowings.[7]

By issuing loan certificates, the clearing house could buy bank assets and economize on the use of cash in the clearing process (where the certificates were accepted as cash) so that cash could be handed out to depositors. Later, clearing house loan certificates were issued directly to the public (see Gorton (1984)). Also, certified checks circulated as cash, and banks accepted them as cash in the clearing process. Certified checks are not dependent on any single account. Further, the checks were stamped "Only Payable Through the Clearing House." This meant that they were the joint liability of the clearing house, rather than of a single bank. These checks circulated as a hand-to-hand currency.

Importantly, by agreeing that certified checks were acceptable as money, the clearing house created a market in these checks. The currency premium on checks was reported in newspapers. The currency premium was the excess check

5. The New York clearing house members were the largest banks in the country and held most of the banking system's reserves, so the solvency of the New York Clearing House was effectively the solvency of the banking system.

6. Clearing house loan certificates were not permanent. They would all be retired at the end of the crisis.

7. This was later the underpinning of the Federal Reserve's discount window when the central bank was established in 1914. Discount window borrowing was to be kept secret.

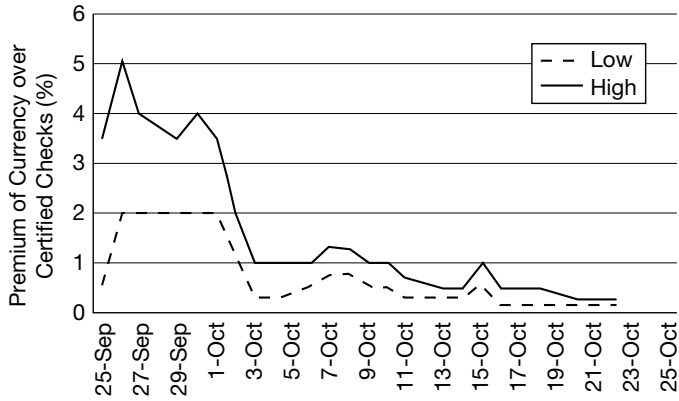


Figure 5.5 Currency Premium for Panic of 1873
SOURCE: Commercial & Financial Chronicle.

amount that needed to be paid to receive a dollar of cash. In other words, an informative financial market was created where the risk of the clearing house failing was priced. Figure 5.5 shows the high and low currency premiums during the Panic of 1873. The high was five percent, meaning that $1.05 of certified checks was needed to buy a dollar of cash. The figure shows the decline in the currency premium, which led to the end of the crisis. If information suggested the clearing house was solvent, it would be impounded in the price so the premium would go down. When there was no longer a premium, convertibility of checks into currency would be resumed.

Also, during the crisis the clearing house would send special examination teams to study the situation of certain specific banks that were the subject of rumors. In the case of a special examination during suspension, the results of the examination were publicized with a certificate of financial health issued by the Clearing House Committee. This occurred even if privately the Clearing House Committee had reservations about the bank's solvency. The certificates issued by the clearing house simply stated that the specific bank was solvent; no detailed information was released. In fact, the detailed results of clearing house examinations were never made public, even in normal times, although bailouts of member banks were public. There were thirteen special examinations conducted during the five panics of the National Banking Era, 1863–1914. See Gorton and Talman (2013).

During the National Banking Era, the New York Clearing House had around sixty members, the largest banks in the country. In the five major panics (1873, 1884, 1890, 1893, 1907) a total of five members failed.

A clearing house, that is a clearing system, is an inherent part of the use of checks as money. It replaced the bank note market. Bank stocks stopped trading also. There was no information leakage and so checks traded at their face value.

During a bank run, the clearing house first suppressed all bank-specific information, so individual banks would no face runs. But, a market did open, a market for claims on the clearing house, and the prices of these claims revealed information about the banking system.

The opacity of banks was endogenously created so that checks could trade at par. The vulnerability to runs meant that the clearing house had to take a central bank-like role as lender-of-last-resort. This role entailed managing the information environment during the crisis. This delicate task meant preventing the revelation of some information while producing and revealing other information. Notably, "transparency" was never the goal, nor would it have been advisable.

5.5. THE FINANCIAL CRISIS OF 2007–2008

The response of the central bank and the government to the financial crisis of 2007–2008 was reminiscent of previous responses in the 19th century. The crisis was an information event, and the information environment had to be managed, most particularly by suppressing some information, hiding some information, and producing and announcing other information.

With the advent of deposit insurance in 1934, depositors had no incentives to produce information about their banks. Deposit insurance guarantees opacity in the sense that no market participants writing checks need worry about the risk of the bank issuing the deposits. Checks are accepted without a second's thought (about the bank). The government's bank examiners look after the banks and the results of their examinations are kept secret.

But, banking systems evolve and so do the forms of bank money. Market economies are not static, just as checks came to replace private bank notes, new forms of bank money have appeared and grown. Indeed, the issue of bank opacity has arisen again with the growth of the "shadow banking system," which refers to a large market in which uninsured short-term bank debt plays a role similar to demand deposits, except that the depositors are large institutions.[8]

In the modern era bank money has expanded to include sale and repurchase agreements ("repo") and asset-backed commercial paper (ABCP). These forms of bank debt are issued by financial intermediaries that were not regulated as commercial banks. ABCP was issued by special purpose vehicles that used the proceeds of issuing the paper to buy asset-backed securities (ABS), bonds backed by portfolios of loans, mortgages, auto loans, credit card receivables, etc. See Gorton and Metrick (2011) and Gorton and Souleles (2006).

8. I do not review all the details here. See Gorton (2010, 2012) and Gorton and Metrick (2012) for details on the crisis.

Similarly, sale and repurchase agreements ("repo") often also used ABS as collateral.

Repo is a form of bank money, which has grown to rival the size of the market for demand deposits. This growth is not surprising because the world has changed. In particular, institutional investors, pension funds, asset managers, sovereign wealth funds now dominate global financial markets. Large nonfinancial firms hold enormous amounts of cash. For these entities there is no insured checking account large enough to accommodate the size of their desired deposits. To meet their demand for a short-term store of value that is safe and that earns interest, the repo market grew enormously.

In a repo, a depositor (lender) deposits money with a dealer bank, usually overnight, and receives interest. To ensure the safety of these deposits, the loan is backed by collateral in the form of bonds (with a market value equal to the amount lent). The depositor takes possession of the bonds. The bonds might be U.S. Treasury bonds, but before the crisis there was a shortage of this form of collateral and much of repo was backed by privately-produced debt, namely, asset-backed securities.

In traditional banking, the bank offers, say, three percent interest to depositors and lends the money to a home buyer at six percent. The bank earns the spread, six minus three. In repo, the depositor is offered three percent, and takes collateral, of say a mortgage-backed security which earns six percent. The six percent goes to the bank, so the bank earns the same spread of six minus three. Furthermore, you can see that the two systems are intimately related because the traditional bank no longer holds the mortgage on its balance sheet. It ends up being financed in the capital markets when it is securitized, that is turned into a bond which is backed by a portfolio of mortgages. So, repo ends up financing traditional bank loans. The point is that "shadow banking" is genuine banking, not some aberration.

ABCP is similar to repo. ABCP is often one to four day maturity and repo is usually overnight. These liabilities serve as a kind of money for large institutions. The short maturity is essential so that depositors have flexible access to their cash. In order for this to function as money the backing collateral must be opaque, as discussed above. For this purpose ABS are ideal. As explained by Gorton and Metrick (2011) and Gorton and Souleles (2006), ABS consist of layers of bonds ordered by seniority (called "tranches") linked to the same large portfolios of loans. The loan portfolios are homogeneous, for example all auto loans or all prime mortgages. Asset classes are never mixed in a portfolio. Also, and importantly, ABS have no traded equity. That part of a transaction is held by the originator. ABS are complicated, opaque, and it is not profitable in normal times to bother doing credit analysis on them. Since ABS have no traded equity, no information is revealed.

Opacity is what makes asset-backed securities ideal for the collateral backing asset-backed commercial paper and repo. Indeed, shadow banking consists of repo and ABCP backed by bonds linked to portfolios of loans. This is real banking: loans are financed by deposits (repo), so to speak, of institutional investors who have a demand for this kind of interest-earning, short-term, saving. This bank money works because the ABS is opaque. But, like the older banking system, which finances loans via deposits, shadow banking is vulnerable to runs, just as the older banking system was prior to deposit insurance.

5.5.1. What Happened?

The financial crisis of 2007–2008 was a bank run on repo and ABCP. Depositors began to worry about the ABS backing their loans and refused to renew their loans. Unlike the bank runs of the nineteenth century and the Great Depression, this run was not visible unless you were on a trading floor. In a run the banks must raise cash. No one will lend to them in the crisis and so they end up having to sell securities, causing bond prices to plummet. But, bond markets—ABS markets—are over-the-counter, and so the plummeting prices were not seen either. Outsider observers saw the effects of the run, namely large banks could not raise enough money and verged on insolvency. Federal Reserve Chairman Ben Bernanke, in his Financial Crisis Inquiry Commission testimony, noted that of the "13 . . . most important financial institutions in the United States, 12 were at the risk of failure within a period of a week or two" (Bernanke (2010)).

All market economies have faced the problem of bank runs, although some countries have avoided the problem for long periods of time. When there is bank run, every society and government has found a way to keep from liquidating its banking system. There have been bailouts, nationalizations, blanket guarantees, and so on. All of these mechanisms have been at root a way to make the suspicions of depositors go away. The government, by virtue of its taxing power, can (usually) do what the private sector cannot, namely, eliminate suspicions about the collateral backing private money.

5.5.2. Overcoming Stigma

I now focus on some aspects of the recent financial crisis related to the information environment. I highlight the information issues faced by the government and show that the goal was to suppress information.

The first issue concerns "stigma." Stigma refers to the negative effects on a bank of information leakage about the bank's borrowing from the discount window. This is perceived to be a sign of weakness, potentially leading to a run on that bank. Fed Chairman Bernanke (2010b):

> Many banks . . . were evidently concerned that if they borrowed from the discount window, and that fact somehow became known to market participants, they would be perceived as weak and, consequently, might come under further pressure from creditors. To address this so-called stigma problem, the Federal Reserve created a new discount window program, the Term Auction Facility (TAF). (p. 2)

Stigma has historically been a problem. In the pre-Federal Reserve period the clearing houses kept secret how many loan certificates each member borrowed. And this secrecy was continued by the Federal Reserve with respect to the discount window. Only that information leaks out so banks are reluctant to use the discount window.

To overcome this problem during the crisis the Federal Reserve designed special lending programs that were based on auctions. The Term Auction Facility (TAF) and other programs obscured which banks were trying to borrow by keeping secret which banks were bidding, how much they were bidding, how much they wanted and which banks got funds. This information was kept secret and since the auction was a coordination mechanism, getting a large number of banks to come to borrow at once, no single bank was stigmatized as weak. Armantier, Ghysels, Sarkar, and Shrader (2011) studied TAF and found that "banks were willing to pay an average premium of at least 37 basis points (and 150 basis points after Lehman's bankruptcy) to borrow from the Term Auction Facility rather than from the discount window."

Also, lending to institutions through the Troubled Asset Relief Program (TARP) was also kept secret.[9] The special lending programs set up by the Fed during the financial crisis, like the clearing house loan certificates, required secrecy so that individual banks would not be singled out by the market.

5.5.3. Banning Short Sales of Bank Stocks

But, wouldn't stock market prices reveal which banks were weak? Yes, the market did reveal which banks were weaker, but not how weak. See Peristiani, Morgan, and Savino (2010). The Federal Reserve undertook "stress tests" to determine

9. In October 2008, the Emergency Economic Stabilization Act of 2008 (Division A of Public Law 110–343) established the Troubled Asset Relief Program (TARP) for the purpose of enabling the Treasury to purchase and guarantee of "troubled assets."

how much capital was needed by each bank. The stress tests (Supervisory Capital Assessment Program, SCAP) were introduced in February 2009. Ten of the 19 largest bank holding companies that underwent the SCAP were required to raise equity capital—by $75 billion in total. Peristiani, Morgan, and Savino (2010) studied the market response to the announcement; it was positive for banks that were required to raise equity. There was no stock price response (abnormal return) for banks that were not required to raise equity.

The SCAP was the only instance where the Federal Reserve produced information and announced it during the crisis. But, the Fed only announced how much capital each bank would need. SCAP was essentially the modern counterpart to the clearing houses' special examinations of members during crises. In both cases, the details of the examinations were not announced. Only a conclusion was announced.

Finally, informative stock prices were viewed as a problem during the financial crisis. In 2008 the U.S. Securities and Exchange Commission (and, in England, the Financial Services Authority) banned short sales of the stock of seventeen large financial firms and also Fannie Mae and Freddie Mac. At the time the SEC (2008) wrote:

> False rumors can lead to a loss of confidence in our markets. Such loss of confidence can lead to panic selling, which may be further exacerbated by "naked" short selling. As a result, the prices of securities may artificially and unnecessarily decline well below the price level that would have resulted from the normal price discovery process. If significant financial institutions are involved, this chain of events can threaten disruption of our markets.

Later, in September 2008, the SEC temporarily prohibited short selling of the stocks of approximately 800 financial firms, required institutional money managers to report short sales and short positions in certain securities, and eased restrictions on the ability of issuers to repurchase their securities.[10]

The short sales bans were attempts to suppress bank-specific information. The academic studies to date show that the short sale bans reduced market liquidity and hindered price discovery, exactly what the bans were intended to do. See, e.g., Beber and Pagano (2013) and the references therein. The academics, however, view short sales bans as misguided. But, in the context of the financial crisis, it appears to have been an attempt to cut off information about specific

10. September Emergency Order Taking Temporary Action to Respond To Market Developments, Exchange Act Release No. 34-58592, 73 Fed. Reg. 55,169 (Sept. 18, 2008), available at http://www.sec.gov/rules/other/2008/34-58592.pdf; Amendment To Emergency Order Taking Temporary Action To Respond To Market Developments, Exchange Act Release No. 58,591A, 73 Fed. Reg. 55,557 (Sept. 21, 2008), available at http://www.sec.gov/rules/other/2008/34-58591a.pdf.

banks, to keep the runs from concentrating on the weak banks. Until the early 1960s bank stocks were already endogenously illiquid and so there could not be short sales. During financial crises then there was no need to ban short sales. In the recent crisis though the information-revealing feature of stock markets were viewed as a problem by the Securities and Exchange Commission. Information that would have revealed weaker banks could have led to runs on those banks. To prevent such runs information was cut off.

Note that to the extent that the short sales bans were successful, investors traded stocks at the wrong prices. Some investors got gains they would not otherwise have gotten, and their trading partners got losses that they would not otherwise have gotten. But, this was—implicitly—viewed as the price for avoiding liquidating the banking system.

5.5.4. Discussion

Did lack of transparency play any role in the crisis? Of course it played a role. That is exactly the vulnerability of banks, they are subject to runs.

When we observe a phenomenon—bank runs—happening over and over again in market economies throughout history, there is a root problem, a common structural problem, an inherent problem. The problem is the vulnerability of bank money. The vulnerability comes from the need for opacity for money to function. Historically, with various forms of money facing runs, the same complaints of complexity and a lack of transparency are heard over and over again.

But, there was another problem too which should not be confused with the opacity that I have been speaking about. Regulators, academics, the media, and the public did not understand how the U.S. financial system had evolved and did not observe the actual runs. As mentioned above, the evolution of the financial system was driven by a number of factors. Over the last thirty years or so there has been the rise of institutional investors and a concurrent decline in the fraction of households that directly hold securities. The fact that regulators, academics, the media, and the public were unaware of the developments in the U.S. banking system and did not see the run is not the same as a "lack of transparency." The inability to see what was going on was not due to a lack of transparency. It is an intellectual problem. It was a failure to understand the evolution of the financial system and a failure to understand the vulnerability of bank money. A lack of understanding of financial history and bank money is at the root of this failure.[11] It was not knowing where to look or, indeed, realizing that it was worth looking

11. I discuss the reasons for this failure in Gorton (2012).

at all. It was simply assumed that the U.S. would never experience a systemic crisis again.

There is clearly a measurement problem. Our forms of measurement, National Income Accounting, the Federal Reserve's Flow of Funds data set, Generally Accepted Accounting Practice, bank Call Report data, and so on, are all important but incomplete in a world with derivative securities and off-balance sheet vehicles. This problem requires augmenting these systems with a system of national risk and liquidity accounting, as proposed by Brunnermeier, Gorton, and Krishnamurthy (2011, 2012) and Bai, Krishnamurthy, and Weymuller (2013).

5.5.5. Summary

The desirability of opacity in banking does not mean that no information should be produced. Banks need to be transparent to the regulators, but that information is kept confidential. This puts the burden on the clearing house and later on bank regulators. Opacity can create systemic risk. But, in the modern era systemic risk is created when regulators are unaware of what information they should be producing. It is their job to distinguish good banks from bad banks. As financial systems evolve, it is important to keep up with this evolution.

5.6. CONCLUSION

Banks are inherently opaque so that their debt can be used as money. This opacity notably developed during the 19th century; it entailed shutting informative markets for bank liabilities (bank notes and bank stock), internalizing that information into the clearing house, which kept the information secret. This is not unlike the modern era in which bank examinations are confidential to the government, and discount window borrowing from the Fed is supposed to be secret.

During financial crises bank coalitions (clearing houses) and central banks have always carefully managed the bank information environment. During crises policies have been aimed at preventing bank runs on individual banks, based on information about specific banks. The financial system can unravel serially if banks are sequentially run on. In general, bank-specific information is suppressed thereby forcing attention to the question of the solvency of the entire banking system. In the 19th century an explicit market pricing the risk of the clearing house being insolvent opened, and when the currency premium went to zero, normalcy returned.

Recently, the problem of bank runs emerged again. The development of new forms of bank money, repo and asset-backed commercial paper, have also been created to be opaque, by being backed by ABS, which itself has no information leakage. The same problems as in the 19th century have reemerged, and the Federal Reserve and the government have rediscovered the modern equivalents, overcoming stigma, introducing stress tests, and trying to suppress information-revealing markets.

REFERENCES

Armantier, Olivier, Eric Ghysels, Asani Sarkar, and Jeffrey Shrader (2011), "Stigma in Financial Markets: Evidence from Liquidity Auctions and Discount Window Borrowing during the Crisis," New York Federal Reserve Bank Staff Report No. 483.

Bai, Jennie, Arvind Krishnamurthy, and Charles-Henri Weymuller (2013), "Measuring Liquidity Mismatch in the Banking Sector," SSRN working paper.

Bartlett, Robert (2012), "Making Banks Transparent," *Vanderbilt Law Review* 65(2), 293–385.

Bernanke, Ben (2010a), "Causes of the Recent Financial and Economic Crisis," Statement by Ben S. Bernanke, Chairman, Board of Governors of the Federal Reserve System, before the Financial Crisis Inquiry Commission, Washington D.C. (September 2, 2010); see http://www.federalreserve.gov/newsevents/testimony/bernanke20100902a.htm.

Bernanke, Ben (2010b), Statement by Ben S. Bernanke, Chairman, Board of Governors of the Federal Reserve System, prepared for the Committee on Financial Services, U.S. House of Representatives, February 10, 2010.

Bolles, Albert (1903), *Practical Banking*, Eleventh Edition (Levey Bro's & Co.).

Brunnermeier, Markus, Gary B. Gorton, and Arvind Krishnamurthy (2012), "Liquidity Mismatch Measurement," (2012), chapter in *Risk Topography: Systemic Risk and Macro Modeling*, edited by Markus Brunnermeier and Arvind Krishnamurthy, forthcoming.

Brunnermeier, Markus, Gary B. Gorton, and Arvind Krishnamurthy (2011), "Risk Topography," National Bureau of Economic Research *Macroeconomics Annual* (Chicago: University of Chicago Press).

Cannon, James (1910), *Clearinghouses* (Washington).

Coase, Ronald (1937), "The Nature of the Firm," *Economica* 4, 386–405.

Dang, Tri Vi, Gary B. Gorton, and Bengt Holmström (2012), "Ignorance and the Optimality of Debt," Working paper, Yale and MIT.

Dang, Tri Vi, Gary B. Gorton, Bengt Holmström, and Guillermo Ordonez (2013), "Banks as Secret Keepers," Working paper.

Dillistin, William (1949), *Bank Note Reporters and Counterfeit Detectors, 1826–1866*, Numismatic Notes and Monographs 114 (New York).

Gibbons, J. S. (1859), *The Banks of New York, Their Dealers, The Clearinghouse, and the Panic of 1857* (New York, 1968; reprint of 1859 original).

Goetzmann, William, Roger Ibbotson, and Liang Peng (2001), "A New Historical Database for the NYSE 1815–1925: Performance and Predictability," *Journal of Financial Markets*, 1–32.

Gorton, Gary B. (1984), "Private Bank Clearinghouses and the Origins of Central Banking," *Business Review—Federal Reserve Bank of Philadelphia*, January/February, 3–12.

Gorton, Gary B. (1985a), "Clearinghouses and the Origin of Central Banking in the United States," *Journal of Economic History* 45, 277–83.

Gorton, Gary B. (1985b), "Bank Suspension of Convertibility," *Journal of Monetary Economics* 15(2) (March 1985): 177–93.

Gorton, Gary B. (1988), "Banking Panics and Business Cycles," *Oxford Economic Papers* 40(4), 751–81.

Gorton, Gary B. (1989), "An Introduction to Van Court's Bank Note Reporter and Counterfeit Detector," data appendix.

Gorton, Gary B. (1996), "Reputation Formation in Early Bank Note Markets," *Journal of Political Economy* 104, 346–97.

Gorton, Gary B. (1999), "Pricing Free Bank Notes," *Journal of Monetary Economics* 44, 33–64.

Gorton, Gary B. (2010), *Slapped by the Invisible Hand: The Panic of 2007* (Oxford University Press).

Gorton, Gary B. (2012), *Misunderstanding Financial Crises* (Oxford University Press).

Gorton, Gary B., and Lixin Huang (2006), "Banking Panics and Endogenous Coalition Formation," *Journal of Monetary Economics* 53, 1613–29.

Gorton, Gary B., and Andrew Metrick (2013), "The Federal Reserve and Panic Prevention: The Role of Financial Regulation and Lender of Last Resort," *Journal of Economic Perspectives*, forthcoming.

Gorton, Gary B., and Andrew Metrick (2012), "Securitized Banking and the Run on Repo," *Journal of Financial Economics* 104, 425–51.

Gorton, Gary B., and Andrew Metrick (2012), "Securitization," chapter in the *Handbook of the Economics of Finance*, volume 2, edited by George Constantinides, Milton Harris, and René Stulz (Elsevier).

Gorton, Gary B., and Don Mullineaux (1987), "The Joint Production of Confidence: Endogenous Regulation and Nineteenth Century Commercial Bank Clearinghouses," Journal of Money, Credit, and Banking 19, 458–68.

Gorton, Gary B., and George Pennacchi (1993), "Financial Intermediaries and Liquidity Creation," *Journal of Finance* 45, 49–72.

Gorton, Gary B., and with Nicholas S. Souleles (2006), "Special Purpose Vehicles and Securitization," chapter in *The Risks of Financial Institutions,* edited by Rene Stulz and Mark Carey (University of Chicago Press).

Gorton, Gary B., and Ellis Talman (2013), "How Did Banking Panics End?," forthcoming working paper.

Holmström, Bengt (1999), "The Firm as a Subeconomy," *Journal of Law, Economics & Organization* 15, 74–102.

Holmström, Bengt (2008), "Discussion of 'The Panic of 2007,' by Gary B. Gorton," In *Maintaining Stability in a Changing Financial System*, Proceedings of the 2008 Jackson Hole Conference, Federal Reserve Bank of Kansas City.

Holmström, Bengt (2011), "The Nature of Liquidity Provision: When Ignorance is Bliss," Presidential Address, Econometric Society, ASSA meetings, Chicago, January 5–8, 2012.

Loeser, John (1940), *The Over-the-Counter Securities Market* (New York: National Quotation Bureau Inc.).

Knox, John J. (1903), *History of Banking in the United States* (New York: B. Rhodes and Company).

Macey, Jonathan, and Geoffrey Miller (1992), "Double Liability of Bank Shareholders: History and Implications," *Wake Forest Law Review* 27, 31–62.

O'Sullivan, Mary (2007), "The Expansion of the U.S. Stock Market, 1885–1930: Historical Facts and Theoretical Fashions," *Enterprise & Society* 8, 489–542.

Peristian, Stavros, Donald Morgan, and Vanessa Savino (2010), "The Information Value of the Stress Test and Bank Opacity," Federal Reserve Bank of New York, Staff Report No. 460.

Redlich, Fritz (1951), *The Molding of American Banking* (New York).

Rockoff, Hugh (1975), *The Free Banking Era: A Reexamination* (New York: Arno).

Rolnick, Arthur, and Warren Weber (1983), "New Evidence on the Free Banking Era," *American Economic Review* 73, 1080–91.

Rolnick, Arthur, and Warren Weber (1984), "The Causes of Free Bank Failures," *Journal of Monetary Economics* 14, 267–91.

Securities and Exchange Commission (2008), July Emergency Order Taking Temporary Action to Respond to Market Developments, Exchange Act Release No. 58,166, 73 Fed. Reg. 42,379 (July 15, 2008), *available at* http://www.sec.gov/rules/other/2008/34-58166.pdf.

Smith, Gordon (1908), "Clearing-House Examinations," *The Bankers' Magazine* LXXVI (1908) (The Bankers Publishing Co.), 177–78.

Stevenson, Charles (1910), "Speculation in Bank Stocks," *Bankers' Magazine* 81, 337–42.

Timberlake, Richard (1984), "The Central Banking Role of Clearinghouse Associations," Journal of Money, Credit and Banking 16, 1–15.

U.S. Department of Commerce (1949), *Historical Statistics of the United States* (U.S. Government Printing Office).

U.S. House of Representatives (1964), "The Market for Bank Stock," Subcommittee on Domestic finance, Committee on Banking and Currency, 88th Congress, 2d Session (December 22, 1964) (Washington D.C.: U.S. Government Printing Office).

Williamson, Oliver (1975), *Markets and Hierarchies: Analysis and Antitrust Implications* (New York: Free Press).

PART II

Banking Panics

6

Bank Suspension of Convertibility

GARY B. GORTON*

6.1. INTRODUCTION

During the nineteenth and early twentieth centuries the American banking system suspended convertibility eight times.[1] That is, during these episodes banks refused to exchange currency for demand deposits upon demand.[2] In each case, suspension was the response to a banking panic which was coincident (or nearly so) with a business cycle downturn [see Cagan (1965) and Gorton (1984)]. A curious aspect of suspension is that despite its explicit illegality, neither banks, depositors, nor the courts opposed it at any time. This paper argues that such accommodating behavior arose because suspension was part of a mutually beneficial arrangement.

* The comments and suggestions of Costas Azariadis, Robert Barro, Bob Defina, Peter Garber, Robert King, Don Mullineaux, Alan Stockman, and Mike Toman are gratefully acknowledged. Errors remain my own. This paper was completed while the author was at the Federal Reserve Bank of Philadelphia. The views expressed in this paper are not necessarily those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

1. Those eight times were: August 1814, Fall 1819, May 1837, October 1839, October 1857, September 1873, July 1893, October 1907. Major panics occurred in all these cases, though during this period suspension also happened without a banking panic (in the 1860's). Also panics happened without suspension. There were also some minor cases of suspension. Friedman and Schwartz (1963) distinguish between the terms "restriction" and "suspension", reserving the latter for the Great Depression during which the government closed banks. Previous episodes were marked by banks "restricting" convertibility between deposits and cash, but unlike the 1933 episode, they carried on all other activities. The analysis here does not involve a government, and no distinction is made between the two terms, though the older usage of "suspension" is maintained throughout, as in Hammond (1957).

2. This refusal was usually qualified in various ways. Banks sometimes limited the amounts of the exchange, or only paid out currency needed for wage bills. For details see Sprague (1910).

The strategy of analysis is to first examine the relations between banks and depositors under full information so that decision rules and outcomes have a useful basis for comparison when an incomplete information setting is subsequently examined. The focus is on the conditions under which suspension of convertibility would be a Pareto-improving part of an assumed demand deposit contract.[3]

With full information there is no role for suspension of convertibility. In the full information setting a banking panic occurs when depositors decide to withdraw all their deposits from banks because of expected capital losses. The expectation of future capital losses is rational and depositors would never agree to suspension because it would prevent them from achieving their optimal portfolio allocations.

Under incomplete information there is a role for suspension. Incomplete information means that depositors do not know the state of banks' investments, but use a noisy indicator to form rational expectations of deposit return rates. A banking panic can be triggered by a movement of the indicator, causing depositors to withdraw all their deposits because of fears of capital losses. By suspending convertibility, banks can signal to depositors that continuation of the investments is mutually beneficial. Suspension, however, only occurs when depositors panic because of expectations formed conditional on observing the noisy indicator, but would not panic if they had full information. Thus, the full information world can be approximated by including suspension as part of the demand deposit contract.

6.2. THE BANKING SYSTEM

The model economy lasts for three periods. Depositors maximize the utility of consumption the first two periods and end of world wealth during the third period. Depositors are retired during the third period and live off their savings. Each depositor begins the world with an inherited endowment of wealth, M_0. Currency and demand deposits are the only available stores of value. The banking system has two essential, exogenously imposed, features. First, individual banks, in a competitive banking system, finance two-period investments (at the beginning of the world) with debt (deposits) and equity. Debt is the senior claimant on a bank's returns. The return on debt may include capital losses, but deposits cannot incur capital gains. Second, depositors may withdraw their deposits at the end of the first period. These two features create the possibility of depositors ending the investment process after the first period.

3. Santomero (1983, sec. I) surveys the literature on why banks exist. Also, see Haubrich and King (1983).

There are two sources of uncertainty in the model. The rate of return to holding currency is random, and the rate of return on banks' investments is random. It is assumed that currency consists of gold coins, and its rate of return is its rate of appreciation or depreciation against goods.[4] At the beginning of the world, the rate of return on currency for period 1 is known, but it is not known what the rate of return on currency will be during period 2. That random variable is realized at the end of period 1.

The rate of return on bank investments is random because of underlying real shocks to produced output upon which banks hold claims. Since the physical realization of the technology that banks have invested in is random, the value of bank investments, referred to as "the state of bank investments", reflects these underlying shocks. Thus, the state of bank investments is a random variable, realized at the ends of periods 1 and 2. Realizations of the state of bank investments determine whether a capital loss is imposed on depositors at the end of each period.

The notation adopted is presented in table 6.1. The rate of return on demand deposits at the end of period 1, the repurchase price, is $(1+r_{d1})(1-\pi_1(\theta_1))$, i.e., the predetermined rate of return on demand deposits (r_{d1}) discounted by the capital loss on demand deposits ($\pi_1(\theta_1)$). The capital loss is determined from the bank's balance sheet, so it follows that

$$(1+r_{d1})[1-\pi_1(\theta_1)] = \min[(1+r_{d1}), \theta_1 L/\Delta]. \qquad (\mathbf{6.1})$$

Whether the depositors incur a capital loss or not depends on the state of bank loans, θ_1, and on the amount that senior claimants can claim,$(1+r_{d1})\Delta$. If a bank cannot repay depositors at the initially agreed upon specie price of deposits, then the deposit price depreciates to reflect the value of the bank's assets. The required depreciation is

$$\begin{aligned} 1-\pi_1(\theta_1) &= 1 && \text{if} \quad \theta_1^* \le \theta_1 \le \overline{\theta}_1, \\ &= \frac{\theta_1 L}{(1+r_{d1})\Delta_1} && \text{if} \quad \underline{\theta}_1 \le \theta_1 \le \theta_1^*, \end{aligned} \qquad (\mathbf{6.2})$$

where the critical value,

$$\theta_1^* = \frac{(1+r_{d1})\Delta_1}{L},$$

just permits satisfaction of the claims against the bank at the fixed price.

Similarly, if depositors hold deposits until the end of the world, then the two-period rate of return on deposits is

4. In general, C can be thought of as an alternative investment which earns λ_1 over the first period, and λ_2 over the second period.

Table 6-1. MODEL NOTATION

r	$\equiv$	the (exogenous) *two-period* rate of return earned on bank investments (%);
r_Q	$\equiv$	the two-period rate of return on equity shares (%);
r_{dt}	$\equiv$	the *one-period* rate of return banks promise to pay on demand deposits over the ith period (%), $i = 1, 2$;
λ_1	$\equiv$	the rate of return on currency over period 1 (net of the services return to deposits) (%);
λ_2	$\equiv$	the rate of return on currency over period 2 (%), a random variable with p.d.f. $g(\lambda_2)$over $\left[\underline{\lambda}_2, \bar{\lambda}_2\right]$;
θ_1	$\equiv$	the state of bank investments at the end of period 1 (%), a random variable with p.d.f. $f(\theta_1)$over $\left[\underline{\theta}_1, \bar{\theta}_1\right]$;
θ_2	$\equiv$	the state of bank investments at the end of the world (%);
$\pi_1(\theta_1)$	$\equiv$	the capital loss on deposits at the end of period 1;
$\pi_2(\theta_2)$	$\equiv$	the capital loss on deposits at the end of the period 2;
C_1	$\equiv$	currency holdings of an individual depositor during period i, $i = 1, 2$;
D_1	$\equiv$	deposit holdings of an individual during period i, $i = 1, 2$;
Q	$\equiv$	the amount of equity at an individual bank;
X_1	$\equiv$	consumption of an individual depositor during period i, $i = 1, 2$;
β	$\equiv$	discount factor;
Δ	$\equiv$	an individual bank's level of debt, the sum of individual depositors' holdings at that bank;
W	$\equiv$	an individual depositor's end of world wealth;
L	$\equiv$	the amount of bank investments at the beginning of period 1;
M_0	$\equiv$	an individual's initial wealth endowment.

$$(1+r_{d1})(1+r_{d2})(1-\pi_2(\theta_2)) = \min\left[(1+r_{d1})(1+r_{d2}), \frac{(1+r)\theta_2 L}{\Delta_2}\right]. \tag{6.3}$$

There is again a critical value, θ_2^*, above which capital losses do not occur, i.e., $\pi_2\left(\theta_2^*\right) = 0$. From (**6.3**), this is

$$\theta_2^* = \frac{(1+r_{d1})(1+r_{d2})\Delta_2}{(1+r)L}.$$

The required depreciation is

$$\begin{aligned} 1-\pi_2(\theta_2) &= 1 && \text{if } \theta_2^* \le \theta_2 \le \bar{\theta}_2, \\ &= \frac{(1+r)\theta_2 L}{(1+r_{d1})(1+r_{d2})\Delta_2} && \text{if } \underline{\theta}_2 \le \theta_2 \le \theta_2^*. \end{aligned} \tag{6.4}$$

Table 6-2. STRUCTURE OF INFORMATION: FULL INFORMATION CASE

	Beginning of Period 1	**Beginning of Period 2**	**End of Period 2**
Known information	$\lambda_1, r_{d1}, r_{d2}$	$\lambda_2, r_{d2}, \theta_1, \pi_1, (\theta_1)$	$\theta_2, \pi_2(\theta_2)$
Depositors' actions	Choose portfolio (C_1, D_1)	Decide to withdraw or not; Choose (C_2, D_2)	(Receive end of world wealth)
Banks' actions	Choose Q, given r_{d1}, r_{d2}, such that $L = Q + \Delta$	If $\underline{\theta}_1 \leq \theta_1 < \theta_1^*$, then set $\pi_1(\theta_1) > 0$	If $\underline{\theta}_2 \leq \theta_2 < \theta_2^*$, then set $\pi_2(\theta_2) > 0$

Faced with these return distributions depositors must choose an initial portfolio at the beginning of period 1 and decide whether to withdraw deposits at the beginning of period 2. These decisions will be based on comparing the prospective returns associated with different portfolios, and will utilize all available information. The information structure of the problem is shown in table 6.2. The information available to depositors at the times described by the first row of table 6.2 is the case of full information (FI). Under full information depositors know the state of bank investments, θ_1, at the beginning of period 2. Expectations are formed rationally, so depositors use θ_1 to compute $\pi_1(\theta_1)$ at the time they are making the decision to withdraw or deposit.

Previously, the states of the bank investments were explained as reflecting real shocks to an underlying production process. If it is assumed that this underlying process exhibits persistence, then the state of bank investments each period is serially correlated. So an observation on the state of bank investments at the end of period 1 allows an inference about what final outcome will be realized at the end of period 2. A specification which incorporates this is

$$\theta_2 - \tilde{\theta}_2 = \gamma(\theta_1 - \tilde{\theta}_1) + \mu, \tag{6.5}$$

where $\gamma > 0$, $E(\theta_1) = \tilde{\theta}_1$, $E(\theta_2) = \tilde{\theta}_2$, $\tilde{\theta}_2 \gg \tilde{\theta}_1$, and μ is white noise with density function $Z(u)$. "E" indicates the expectation operator.

Banks and depositors are assumed to know the process (**6.5**). At the beginning of period 2, having observed θ_1, depositors' expectation of θ_2 is

$$E_1(\theta_2) \equiv E(\theta_2|\theta_1) = \tilde{\theta}_2 + \gamma(\theta_1 - \tilde{\theta}_1).$$

Using eqs. (**6.4**) and (**6.5**), the expected capital loss at the end of period 2, conditional on having observed θ_1 at the end of period 1, is

$$E_1[\pi_2(\theta_2)] = \int_{\underline{\mu}}^{\mu^*} \left\{ 1 - \frac{(1+r)L\left[\tilde{\theta}_2 + \gamma\left(\theta_1 - \tilde{\theta}_1\right) + \mu\right]}{(1+r_{d1})(1+r_{d2})\Delta} \right\} Z(\mu)\, d\mu, \tag{6.6}$$

where

$$\mu^* = \theta_2^* - \tilde{\theta}_2 - \gamma\left(\theta_1 - \tilde{\theta}_1\right).$$

6.3. THE DEPOSITORS' FULL INFORMATION PROBLEM

At the beginning of the world, depositors choose a portfolio to get a consumption path. The representative depositor faces the following problem:

$$\max V_0 = E_0\left\{U(X_1) + \beta U(X_2) + \beta^2 \Lambda(W)\right\}, \qquad \textbf{(I)}$$

subject to

(i) $X_1 + C_1 + D_1 \leq M_0$,

(ii) $X_2 + C_2 \leq (1+\lambda_1)C_1 + (1+r_{d1})\left[1 - \pi_1(\theta_1)\right](D_1 - D_2)$,

(iii) $W = (1+\lambda_2)C_2 + (1+r_{d1})(1+r_{d2})\left[1 - \pi_2(\theta_2)\right]D_2$.

Constraint (ii) requires second-period consumption (X_2) and second-period currency holdings (C_2) to be financed by the value of the depositor's portfolio realized at the end of period 1. Constraint (ii) applies the capital loss on deposits only to the amount of deposits withdrawn at the end of the first period, i.e., $(D_1 - D_2)$. We assume returns are bounded such that $D_2 \leq D_1$, i.e., the representative depositor never deposits more at the end of period 1. Constraint (iii) determines the representative depositor's end of world wealth as a function of returns realized at the end of period 2.

Working backwards in typical dynamic programming fashion, we start by analyzing the problem faced by agents at the end of the first period. That problem is

$$\max V_1 = E_1\{U(X_2) + \beta\Lambda(W)\}, \qquad \textbf{(II)}$$

subject to (ii), (iii).

Assume that depositors are risk-averse with respect to lotteries on consumption during periods 1 and 2, but are risk-neutral with respect to retirement wealth. This assumption simplifies the analysis and focuses attention on the problem of interest. The assumption causes depositors to choose portfolios which are corner solutions; depositors hold either currency or deposits, but not both. Consequently, if depositors hold deposits at the beginning of the world, then all their wealth is in this form. If depositors withdraw their deposits at the end of period 1, they withdraw all their deposits, switching completely to currency. Under this assumption,

$$\Lambda(W) = A + BW, \qquad A, B > 0,$$

and using

$$E_1(W) = (1+\lambda_2)C_2 + (1+r_{d1})(1+r_{d2})[1 - E_1(\pi_2(\theta_2))]D_2,$$

we find that if depositors start the world holding deposits, then they will withdraw *all* their deposits if

$$(1+\lambda_2)[1-\pi_1(\theta_1)] > (1+r_{d2})[1-E_1(\pi_2(\theta_2)]. \qquad (\mathbf{6.7})$$

According to (**6.7**), depositors withdraw their deposits if the known rate of return to currency over period 2 is greater than the expected rate of return to holding deposits over period 2, accounting for the capital loss associated with withdrawing. (λ_2 and θ_1 are independent.) This decision rule for withdrawing, which compares a known return to an expected return, is the result of depositors' risk neutrality toward end of world wealth, and the fact that, knowing θ_1, second-period utility is not uncertain.

For each realized θ_1, there exists a critical value of the rate of return on currency, $\lambda_2^*(\theta_1)$, such that depositors are just indifferent between withdrawing and not withdrawing,

$$[1+\lambda_2^*(\theta_1)] = \frac{(1+r_{d2})[1-E_1(\pi_2(\theta_2)]}{[1-\pi_1(\theta_1)]}. \qquad (\mathbf{6.8})$$

That is, the decision rule is to withdraw if $\lambda_2 > \lambda_2^*(\theta_1)$, which divides the area of possible (λ_2, θ_1) realizations into a region over which depositors will withdraw their deposits and the remainder over which they will not withdraw (see figure 6.1).

The slope of rule (**6.8**) depends on the implications of the θ_1 realization for the prospective return on deposits at the end of period 2,

$$\frac{\partial \lambda_2^*(\theta_1)}{\partial \theta_1} = \frac{(1+r_{d2})\gamma\ \Gamma}{\theta_2^*} \qquad \text{if} \quad \theta_1^* \le \theta_1 \le \bar{\theta},$$

$$= \frac{(1+r)\gamma\ \Gamma}{\theta_2^*} - \frac{(1+r_{d2})[1-E_1(\pi_2(\theta_2))]}{[1-\pi_1(\theta_1)]^2\theta_1^*} \qquad \text{if} \quad \underline{\theta}_1 \le \theta_1 < \theta_1^*,$$

where

$$\Gamma = 1 - \int_{\underline{\mu}}^{\mu^*} Z(\mu)\, d\mu,$$

which is the probability of the banking system not failing at the end of period 2.

The slope of the withdraw rule is positive with respect to increases in θ_1. To see this recall that above we assumed that a low θ_1 realization currently implies a lower θ_2 realization next period since $\gamma > 0$ in (**6.5**). Now consider the range of θ_1 realizations over which there is no capital loss on deposits at the end of period

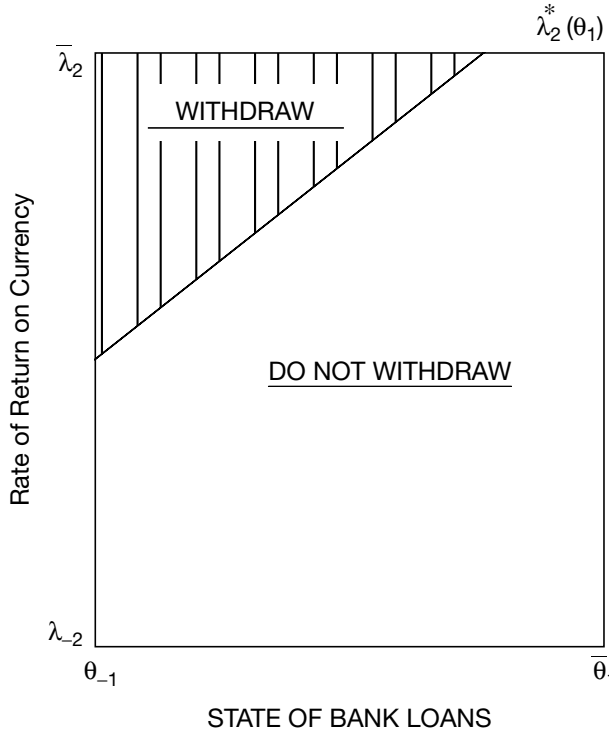


Figure 6.1 The full information withdraw rule

1, i.e., $\theta_1^* \leq \theta_1 \leq \bar{\theta}_1$. Over this range, as θ_1 increases, $E_1(\pi_2(\theta_2))$ decreases, increasing $\lambda_2^*(\theta_1)$ since the prospective return to deposits at the end of the second period is more favorable.

Over the range where there is currently a capital loss on deposits, $\underline{\theta}_1 \leq \theta_1 < \theta_1^*$, two forces pull $\lambda_2^*(\theta_1)$ in opposite directions. As θ_1 increases over this range, both the current and prospective capital losses decline. If γ is large enough, then, as θ_1 increases, $\lambda_2^*(\theta_1)$ increases because $E_1(\pi_2(\theta_2))$ declines by more than $\pi_1(\theta_1)$. Though it is not necessary for what follows, diagrams which follow assume that γ is large enough that the slope over this range is positive.

6.4. DEPOSIT MARKET EQUILIBRIUM UNDER FULL INFORMATION

Banks are risk-neutral and there are no bankruptcy costs. The investment process is assumed to be such that a positive return on equity can only be earned if depositors do not withdraw their deposits at the end of period 1. Banks are required to earn an expected return on equity no greater than an exogenously given number, $\bar{r}_Q$. Once chosen, the level of equity cannot be changed

by the bank at the end of the first period. Given the depositors' rule for withdrawing, any initial level of debt, Δ, and depositors' choice of r_d, a bank then chooses a full information equity level, Q^F, by equating expected profits with the return on equity, $Q\left(1+\bar{r}_Q\right)$. This yields a decision rule for equity (see appendix).

At the beginning of the world each bank announces its rule for choosing an amount of equity. Depositors choose a portfolio at the beginning of the world to maximize expected utility (assuming $r_{d1} = r_{d2} = r_d$) knowing the relations between the expected capital loss, the promised rate of return on deposits (r_d), the total level of deposits at the bank (Δ), and the banks' rules for equity.[5] Since depositors have identical attitudes toward risk and can choose any amount of risk, they distribute themselves across banks so that, in equilibrium, all banks have identical debt–equity ratios $(Q/\Delta)$and deposit rates (r_d).[6]

At the end of the first period, banks and depositors observe λ_2 and θ_1. Depositors re-evaluate their portfolios and decide whether to withdraw their deposits or not. The information in θ_1 about the likely realization of capital losses at the end of period 2 is rationally used by depositors in making the decision to withdraw or not. If depositors withdraw their deposits, then they end the investment process. The decision to withdraw deposits at the end of period 1 is an optimal decision in the presence of full information.[7]

6.5. THE INCOMPLETE INFORMATION EQUILIBRIUM

There is no role for suspension to play under full information conditions. Under full information, depositors know the stochastic process of shocks to bank investments, eq. (**6.5**), and observe the θ_1 realization at the end of period one. Conditional on the observed θ_1, and knowing λ_2, depositors withdraw all their

5. Second-period consumption is implicitly determined by the first-order conditions for (II). Using that function and the withdraw rule, eq. (6.8), the depositors' first-period problem way be solved. Appendix A of Gorton (1982) solves the depositors' first-period problem. At the beginning of the world depositors choose an initial portfolio, (C_1, D_1), and r_d. Gorton (1982) also considers indexing r_{d2} by θ_1.

6. This is a result of constraining depositors to each have only one bank, i.e., an underlying assumption about returns to scale in the transaction technology. The results do not depend on identical debt–equity ratios in equilibrium.

7. At the end of the first period, if depositors decide to withdraw and $\pi_1(\theta_1) > 0$, then there is the possibility of renegotiation of the contract. This possibility is considered in section VI of Gorton (1982). The initial contract could also incorporate this possibility by indexing r_{d2} by θ_1 and λ_2. This would change the area over which the bank would be declared bankrupt, but under incomplete information, does not eliminate suspension as a Pareto-improving part of the contract.

deposits at the end of period 1 when they expect large enough capital losses on deposits at the end of period 2 as determined by the withdraw rule. In this case it is optimal for depositors to withdraw their deposits, and suspension would be a constraint preventing the realization of that decision.

Suspension of convertibility, however, can play a role if depositors are incompletely informed about the state of bank investments. The incomplete information setting is assumed here, but has recently been rationalized by several researchers [e.g., Boyd and Prescott (1984)]. Without full information depositors make mistakes *relative to full information.* It is the existence of these potential mistakes which creates the possibility of a signalling role for suspension, that is, suspension by banks can signal to depositors that they have made a suboptimal decision relative to full information.

Suppose that depositors do not know θ_1 at the end of period 1, but banks observe θ_1. Without knowledge of θ_1, depositors cannot compute $\pi_1(\theta_1)$ exactly. Nor can depositors revise their expectation of $\pi_2(\theta_2)$. Depositors, however, will be assumed to have a noisy indicator of θ_1. For purposes of the model it is convenient to let λ_2 serve as the indicator of the value of banks' portfolios. Suppose that λ_2 is negatively correlated with θ_1 and that depositors observe λ_2 at the end of period 1. The assumed correlation means that gold coins appreciate during "bad" times, i.e., when θ_1 is low.

Again working backwards, at the end of period 1, depositors, with incomplete information, maximize expected second-period utility conditional on having observed λ_2,

$$\max V_2 = E[U(X_2)\,|\lambda_2] + \beta E[\Lambda(W)\,|\lambda_2], \qquad \textbf{(III)}$$

subject to (ii) and (iii).

(Expectations conditional on having observed λ_2 are indicated by '$|\lambda_2$'.) As before, depositors will behave as "plungers" and hold either all currency or all deposits over period 2. Under incomplete information, depositors decide to withdraw if $\lambda_2 > \lambda_2^{**}$, where λ_2 is observed and λ_2^{**} is given by

$$\left(1+\lambda_2^{**}\right) E\left\{[1-\pi_1(\theta_1)]\,U'_{x_2}\right\} = (1+r_{d2})\,[1-E[\pi_2(\theta_2)]]\,E\left[U'_{x_2}\right]. \qquad \textbf{(6.9)}$$

The expectations in (**6.9**) are conditional on having observed λ_2.

Under full information, $\lambda_2^*(\theta_1)$ was chosen to equate the marginal utility of withdrawing with the marginal utility of not withdrawing. Now, λ_2^{**} is chosen to equate the *expected* marginal utilities of withdrawing and not withdrawing. Since θ_1 is not known, under incomplete information, second-period utility is uncertain, so expected marginal utilities (conditional on having observed λ_2) enter the decision rule for withdrawing.

Given depositors' decision rule for withdrawing, banks choose a different rule for their choice of equity (see appendix). Then given the rule for withdrawing

and the bank's rule for choosing equity, depositors, at the beginning of the world, choose a level of deposits and an initial r_d. In general, under incomplete information, Q, Δ, and r_d.will be chosen differently, so that all the variables depending on these, $\theta_1^*, \theta_2^*, \pi_1(\theta_1), \pi_2(\theta_2)$, will have different values under incomplete information.

The full and incomplete information rules for depositor withdrawal are shown in figure 6.2.[8] The incomplete information rule cannot replicate the full information decisions, so depositors are worse off.[9] In particular, a realization of (λ_2, θ_1) in area A or area C results in an incorrect decision by depositors under incomplete information.[10] In area A depositors withdraw all their deposits under incomplete information, when they would not if they had full information. In area C, depositors do *not* withdraw deposits when they would if they had full information. These mistakes result from the fact that the indicator, λ_2, does not reveal the exact state of bank investments.

6.6. THE SUSPENSION CONTRACT

Both banks and depositors would prefer to avoid the banking panic occurring in area A. Depositors prefer to avoid the area A mistake because withdrawing in area A reduces expected end of world wealth. Banks prefer that the investment process not be ended so that a (positive expected) return on equity can be earned. The situation, however, is asymmetric because only depositors have an incentive to avoid area C. A mistake by depositors in area C is to the advantage of banks since depositors do not end the investment process (which they would if they

8. Comparing the banks' decision rules under full and incomplete information (see appendix) it is apparent that if depositors choose λ_2^{**} such that areas A and C are equal (see figure 6.2), then Δ, Q, and r_d would be the same under either information assumption. This, however, cannot be the solution under incomplete information. Under incomplete information depositors will choose some combination of a lower level of deposits and a lower r_d . In that case the expected marginal value of the withdraw option [see appendix A of Gorton (1982)] under full information would be higher than it would be under incomplete information by exactly the marginal utility over areas A and C, which, moreover, would be equal (i.e., $A = C$). In this case, however, depositors' beginning of the world first-order condition, eq. (A9) of appendix A of Gorton (1982), cannot-possibly be satisfied. Satisfying it requires lowering D and r_d , which would lower λ_2^{**}, so that area C would be less than area A.

9. Since closed form solutions for the beginning of the world problems cannot be obtained [see appendix A of Gorton (1982)], it cannot be proven that areas A and C, in the figure, exist. In what follows it is assumed that, under incomplete information, λ_2^{**} is chosen such that areas A and C exist.

10. Since depositors do not observe θ_1, but observe λ_2 and form a conditional expectation of θ_1 using λ_2 , figure 6.2 has only one relevant dimension under incomplete information. It is drawn in two dimensions for illustrative purposes.

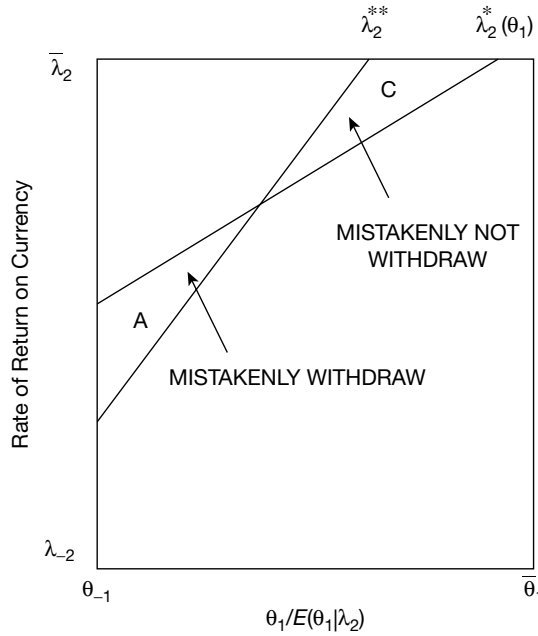


Figure 6.2 Full and incomplete information rules

had full information). A prestate agreement which avoided the effects of the area *A* banking panic would be mutually beneficial to both banks and depositors. But this would not be the case for area *C*.

Since banks and depositors are asymmetrically informed we modify the model to allow information about the state of banks, θ_1, to be transmitted to depositors at a cost. Any realization of θ_1 is known only by banks unless a verification cost is borne. [See Townsend (1979).] In this setting we will consider a prestate agreement which states when verification is to take place and what the outcome of exchange is to be, contingent on the state revealed. If banks signal when verification is to take place, submit to verification, and abide by the prestate specified outcome, then the contract is said to be incentive compatible.

The only difficulty is the asymmetry between areas *A* and *C*. Both banks and depositors have incentives to avoid area *A*, but only depositors want to avoid area *C*. However, if the prestate agreement refers only to area *A*, allowing this mistake to be avoided, then depositors will be compensated for the area *C* mistake. Since the expected rate of return on equity cannot exceed $\bar{r}_Q$, the gain to banks from avoiding area *A* will accrue to depositors.

Consider the following arrangement between a bank and its depositors. If depositors, under incomplete information, withdraw their deposits at the end of period 1 because $\lambda_2 > \lambda_2^{**}$, then the bank is allowed to suspend convertibility if it chooses. Suspension, however, requires the equity holders of the bank to pay

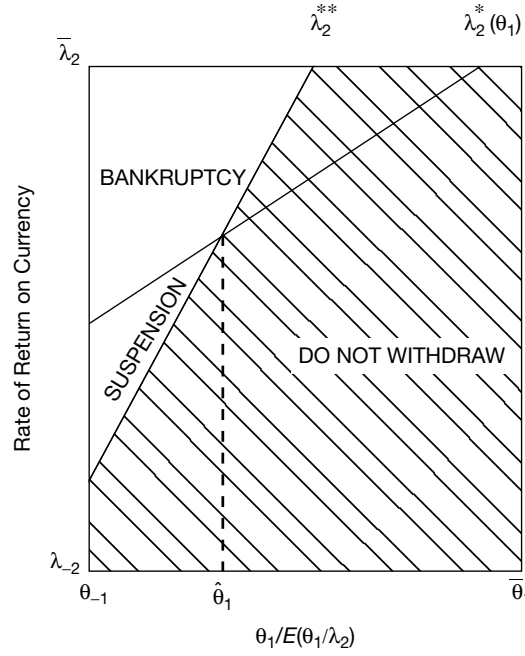


Figure 6.3 The suspension contract

a verification cost proportional to its debt, $v\Delta$. If the verification cost is paid, then the true realization of θ_1 is determined and revealed to depositors.

There is no incentive for a bank to suspend outside area *A*. After verification, depositors will demand the return of their deposits anyway and the verification cost would be unnecessarily lost to equity holders. If depositors withdraw and there is no capital loss, i.e., $\theta_1^* \leq \theta_1 \leq \bar{\theta}_1$, then the bank can pay off the claims of depositors without inflicting capital losses. But the bank has an incentive to suspend and leave the investments undisturbed. However, suspension would require verification, so that depositors would receive $(1 + r_{d1})\Delta$ and equity holders would be liable for $v\Delta$. This strategy cannot be optimal. The situation is similar if depositors withdraw and there is a positive capital loss, i.e., $\underline{\theta}_1 \leq \theta_1 \leq \theta_1^*$. While the bank has an incentive to suspend, verification would show that, unless the realization was in area *A*, depositors would demand their deposits and equity holders would have to pay the verification cost $(v\Delta)$. The bank, therefore, only suspends in area *A*. Therefore, this agreement is incentive compatible.

The contract is pictured in figure 6.3. Since depositors now withdraw over a smaller area of the space of possible first-period realizations, the expected rate of return on an equity share, $E(r_Q)$, will exceed $\bar{r}_Q$ when $E(r_Q)$ is computed using the banks' decision rule for equity under incomplete information (see appendix). Under the suspension contract the amount of equity chosen will exceed the amount chosen without suspension in the contract (see appendix).

In other words, to satisfy the constraint banks are forced to return the increase in expected profits to depositors. Banks raise their equity–debt ratios, making deposits "safer" by reducing expected capital losses. This is the source of the welfare gain to depositors. But as long as $\nu > 0$, depositors cannot achieve the level of expected utility attainable under full information. (See appendix.)

During a banking panic suspension signals that the realization is in area A. The verification process accompanying suspension allows depositors to determine the state of bank investments, information not fully revealed by λ_2. In effect, depositors only monitor (or monitor more intensely) when they have reason to expect that high capital losses on deposits are more probable, i.e., "high" realizations of λ_2. The Pareto-improvement captured by the suspension contract originates in avoiding the results of the panic which would occur without suspension.

6.7. CONCLUSION

The view of panic and suspension presented here may best be described as an information-based explanation. Without full information about the state of bank investments, a panic can be rationally triggered by movements in a noisy indicator of the state of bank investments. The panic is "rational" because the indicator contains useful information; it is, in fact, correlated with the state of bank investments. The indicator is not an intrinsically irrelevant variable. If a panic occurs, banks, with superior information, can signal to depositors that continuation of the investment process is mutually beneficial. Suspension circumvents the realization of suboptimal depositor withdrawals which are based on (rational) fears of capital losses.

The information-based explanation of panic and suspension implies that these events are predictable on the basis of prior information. That is, panic and suspension are not random events, but are related to changes in expected returns caused by movements in the indicator. While the indicator used in the model, λ_2, should not be interpreted literally (as the rate of return on currency), the model makes fairly strong predictions about when panics and suspensions should occur. In a study of the National Banking Era (1865–1914), Gorton (1984) uses the liabilities of failed non-financial businesses as the indicator and shows that every time this variable reached a defined critical level there was a panic. Other researchers have cited, as indicators, the failure of particular large, non-financial corporations [e.g., Friedman and Schwartz (1963)], or "seasonal stringency" [e.g., Kemmerer (1910)].

The information-based explanation of panic and suspension contrasts sharply with what may be described as bubble explanations. Recent examples of this latter view include Diamond and Dybvig (1983) and Waldo (1982). In these models the occurrence of an intrinsically irrelevant event can cause a panic

because of the exogenous imposition of a first come, first serve rule for bank payouts to depositors. Hence, individual depositors have an incentive to "beat" runs and anything which happens causing them to anticipate a panic causes the panic. Unfortunately, bubble explanations appear to place no testable restrictions on the data.

APPENDIX

Under full information, given r_d, a level of debt, Δ, and the depositors' rule for withdrawing [eq. (**6.7**) in the text], each bank, at the beginning of the world, chooses an amount of equity, Q^F,by equating expected profits with $\left(1+\bar{r}_Q\right)Q$, where $\bar{r}_Q$ is the maximum rate of return on equity. The solution to the banks' problem is

$$\frac{Q^F}{\Delta}=\frac{\left[E_0\left(1+r\right)|NW\right]-E_0\left[\left(1+r_d\right)^2|NW\right]}{1+\bar{r}_Q-E_0\left[\left(1+r\right)|NW\right]}, \tag{**6A.1**}$$

where

$$E_0[(1+r)|NW]\equiv G(1+r)\int_{\mu^*}^{\overline{\mu}}[\widetilde{\theta}_2+\mu]Z(\mu)d\mu,$$

$$E_0\left[(1+r_d)^2|NW\right]\equiv G(1+r_d)^2\widetilde{\theta}_1\int_{\mu^*}^{\bar{\mu}}Z(\mu)d\mu,$$

$$G\equiv\int_{\underline{\theta}_1}^{\bar{\theta}_1}\int_{\underline{\lambda}_2}^{\lambda_2^{*(\theta_1)}}g(\lambda_2)f(\theta_1)d\lambda_2 d\theta_1.$$

"NW" indicates conditional on not withdrawing. E_0 indicates the expectation at the beginning of the world. Under incomplete information, each bank chooses an amount of equity, Q^I, in the same way except that the depositors' rule for withdrawing is different [eq. (**6.9**) in the text]. The form of the banks' solution is the same as (6A.l), except, under incomplete information,

$$G\equiv\int_{\underline{\theta}_1}^{\bar{\theta}_1}\int_{\underline{\lambda}_2}^{\lambda_2^{**}}g\left(\lambda_2\right)f\left(\theta_1\right)d\lambda_2 d\theta_1.$$

Under the suspension contract, the banks' decision rule is given by

$$\frac{Q^S}{\Delta}=\frac{E_0\left[\left(1+r\right)|NW,II,S\right]-E_0\left[\left(1+r_d\right)^2|NW,II,S\right]+E_0\left[\left(1+r\right)|S\right]}{1+r_Q-E_0\left[\left(1+r\right)|NW,II,S\right]-E_0\left[\left(1+r\right)|S\right]}$$

$$-\frac{E_0\left[\left(1+r_d\right)^2+r|S\right]}{1+r_Q-E_0\left[\left(1+r\right)|NW,II,S\right]-E_0\left[\left(1+r\right)|S\right]}, \tag{**6A.2**}$$

where

$$E_0\left[(1+r)^2\,|NW,II,S\right] \equiv G(1+r)\int_{\mu^*}^{\bar{\mu}}\left[\tilde{\theta}_2+\mu\right]Z(\mu)\,d\mu,$$

$$E_0\left[(1+r_d)^2\,|NW,II,S\right] \equiv G(1+r_d)^2\,\tilde{\theta}_1\int_{\mu^*}^{\bar{\mu}}Z(\mu)\,d\mu,$$

$$E_0\left[(1+r)\,|S\right] \equiv A(1+r)\int_{\mu^*}^{\bar{\mu}}\left[\tilde{\theta}_2+\mu\right]Z(\mu)\,d\mu,$$

$$E_0\left[(1+r_d)^2+v|S\right] \equiv A\left[(1+r_d)^2+v\right]\int_{\mu^*}^{\bar{\mu}}Z(\mu)\,d\mu,$$

$$G \equiv \int_{\underline{\theta}_1}^{\bar{\theta}_1}\int_{\underline{\lambda}_2}^{\lambda_2^{**}} g(\lambda_2)f(\theta_1)\,d\lambda_2 d\theta_1,$$

$$A \equiv \int_{\theta_1}^{\tilde{\theta}_1}\int_{\lambda_2^{**}}^{\lambda_2^{*(\theta_1)}} g(d_2)f(\theta_1)\,d\lambda_2 d\theta_1.$$

"S" indicates that the solution is conditional on suspension being part of the contract; "II" indicates incomplete information. To compare this decision rule for equity to the incomplete information decision rule for equity, suppose depositors chose the same Δ and r_d as under incomplete information. Then eq. (6A.2) can be written as

$$\frac{Q^S}{\Delta} = \frac{Q^I}{\Delta} + \frac{E_0\left[(1+r)\,|S\right](1+\Delta/Q) - E_0\left[(1+r_d)^2+v|S\right]}{1+\bar{r}_Q - E_0\left[(1+r)\,|\mathrm{N}W,II\right]}. \qquad \textbf{(6A.3)}$$

Therefore, if depositors chose the same Δ and r_d, $Q^S > Q^I$.

Under the suspension contract depositors withdraw with suspension allowed if $\lambda_2 > \lambda_2^{**}$, where λ_2^{**} is given by eq. (**6.9**) of the text, but λ_2^{**} is computed given the banks' decision rule (6A.2). Solving eq. (6.10) for the equity–debt ratio, get:

$$\left[1+Q^S/\Delta\right] = \frac{(1+r_d)^2\int_{\underline{\mu}}^{\mu^*}\int_{\underline{\theta}_1}^{\bar{\theta}_1} f(\theta_1|\lambda_2)Z(\mu)\,d\theta_1 d\mu \cdot \int_{\underline{\theta}_1}^{\bar{\theta}_1} U_{x_2} f(\theta_1|\lambda_2)\,d\theta_1}{\left(1+\lambda_2^{**}\right)\int_{\underline{\theta}_1}^{\bar{\theta}_1}\theta_1 U_{x_2} f(\theta_1|\lambda_2)\,d\theta_1}.$$

$$+\frac{(1+r)}{(1+r_d)}\int_{\underline{\mu}}^{\mu^*}\int_{\underline{\theta}_1}^{\bar{\theta}_1}\left[\bar{\theta}_2+\mu\right]f(\theta_1|\lambda_2)\,d\theta_1 \cdot \int_{\underline{\theta}_1}^{\bar{\theta}_1} U_{x_2} f(\theta_1|\lambda_2)\,d\theta_1.$$

If depositors chose the same Δ and r_d under the suspension contract as under incomplete information, then the right side of (6A.4) would be the same in both cases. Then since $Q^S > Q^I$, λ_2^{**} would have to be lower everywhere.

Call this choice of λ_2^{**}, λ_2^{***}. Since deposits are now safer, under the assumption that depositors choose the same Δ and r_d as under incomplete information, $\lambda_2^{***} < \lambda_2^{**}$, increasing the area of suspension, since capital losses decline, and minimizing the error associated with area C. Depositors, however, cannot completely eliminate area C because $v > 0$. Depending on depositors beginning of the world first-order conditions, however, the compensation to depositors can be absorbed by depositing more and raising r_d, which raises λ_2^{***}. The gain to depositors remains, but the form changes.

REFERENCES

Boyd, John H. and Edward C. Prescott, 1984, Financial intermediary-coalitions, Federal Reserve Bank of Minneapolis Research Department working paper no. 250.

Cagan, Phillip, 1965, Determinants and effects of changes in the stock of money, 1875–1960 (National Bureau of Economic Research, New York).

Diamond, Douglas and Philip Dybvig, 1983, Bank runs, deposit insurance, and liquidity, *Journal of Political Economy* 91, no. 3.

Friedman, M. and A. Schwartz, 1963, A monetary history of the United States, 1867–1960 (Princeton University Press, Princeton, NJ).

Gorton, G., 1982, Bank suspension of convertibility. Federal Reserve Bank of Philadelphia mimeo.

Gorton, G., 1984, Banking panics and business cycles, Federal Reserve Bank of Philadelphia mimeo.

Hammond, Bray, 1957, Banks and politics in America (Princeton University Press, Princeton, NJ).

Haubrich, Joseph and Robert King, 1983, Banking and insurance. University of Rochester mimeo.

Kemmerer, Edwin W., 1910, Seasonal variations in the relative demand for money and capital in the United States (Government Printing Office, Washington, DC).

Santomero, Anthony, 1983, Modeling the banking firm, University of Pennsylvania mimeo.

Sprague, O., 1910, History of crises under the national banking system (S. Doc. No. 538, 61st Congress, 2nd Sess.), National Monetary Commission.

Townsend, R., 1979, Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21.

Waldo, Douglas G., 1982, Bank runs and an insured banking system, University of Florida discussion paper no. 61.

7

Banking Panics and Business Cycles*

GARY B. GORTON*

7.1. INTRODUCTION

The nearly universal experience of banking panics has led many governments to regulate the banking industry. Economists, too, have increasingly focused on panics as evidence of bank uniqueness. Yet, competing theories to explain banking panics have never been tested. Are banking panics caused by the same relations governing consumer behavior during nonpanic times? Are panics random events, or are panics associated with movements in expected returns, in particular, with movements in perceived risk which are predictable on the basis of prior information? If so, what is the relevant information? Using newly constructed data this study addresses these questions by examination of the seven panics during the U.S. National Banking Era (1863–1914). Depositor behavior under subsequent monetary regimes is also examined. In all, one hundred years of depositor behavior are examined.

* The comments and assistance of Andy Abel, Robert Barro, Phillip Cagan, Bob DeFina, Mike Dotsey, Mark Edwards, Stanley Engerman, Lauren Feinstone, Claudia Goldin, Jack Guttentag, Robert King, Erv Miller, Jeremy Siegel, Alan Stockman, John Taylor, Steve Zeldes, two anonymous referees, and the University of Pennsylvania Macro Lunch Group, were helpful and greatly appreciated. They are not responsible for errors. The research assistance of Earl Pearsall, Elaine Ross, and Wendy Tann was invaluable for this work, as was the programming assistance of Steve Franklin, and Wells Vinton. Thanks to Robert Avery for help with the Tobit program, CRAWTRAN. This study was initiated while the author was at the Federal Reserve Bank of Philadelphia. The study was completed using the Philadelphia Fed's computers, thanks to Richard Lang. The views expressed in this paper are not necessarily those of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

A common view of panics is that they are random events, perhaps self-confirming equilibria in settings with multiple equilibria, caused by shifts in the beliefs of agents which are unrelated to the real economy. An alternative view makes panics less mysterious. Agents cannot discriminate between the riskiness of various banks because they lack bank-specific information. Aggregate information may then be used to assess risk, in which case it can occur that all banks may be perceived to be riskier. Consumers then withdraw enough to cause a panic. While the former hypothesis is not testable, it suggests that panics are special events and implies that banks are inherently flawed. The latter hypothesis is testable; it suggests that movements in variables predicting deposit riskiness cause panics just as such movements would be used to price such risk at all other times. This hypothesis links panics to occurrences of a threshold value of some variable predicting the riskiness of bank deposits.

The thrust of this paper is to differentiate between these two hypotheses. Since the former hypothesis imposes no restrictions on the data, this will, needless to say, be difficult. I, therefore, focus attention on the second hypothesis. The analysis is conducted along two lines. A reduced-form equation describing the behavior of the deposit-currency ratio is studied, and correlations in the data using only the panic dates are studied.

The results suggest that banking panics can be explained by the economic theory explaining consumer behavior during nonpanic times. Banking panics during the U.S. National Banking Era were systematic responses by depositors to changing perceptions of risk, based on the arrival of new information rather than random events. In fact, I show below that every time a variable predicting a recession reached a threshold level, a panic occurred. All the large movements in this variable exactly correspond to large movements in a consumption-beta-type measure of deposit riskiness. The risk measure also reaches a threshold or critical level at panic dates. Panics did not occur at other times. The interpretation is intuitive. Consumers know that during recessions they will want to dissave, drawing down bank accounts. But, banks, like other firms, tend to fail during recessions. When consumers forecast a coming recession they withdraw deposits in advance to avoid losses due to bank failure.

Thus, the analysis confirms that there is something special about panics, but not in the way suggested by theories of self-fulfilling panics or random shifts of depositor beliefs. Rather, depositor behavior during panics is accurately described by a model which characterizes their behavior at other times. But, the information arriving about a coming recession (while noisy) reaches a critical level; this is "special."

The panics of the 1930s, however, cannot be ascribed to the same pattern of consumer behavior. An estimated counterfactual shows that had the downturn of the thirties come during the National Banking Era, losses to depositors would have been four to five times lower; the number of banks that failed during

the thirties was roughly twenty-five times what it would have been had the pre-Federal Reserve System institutions been in place. The banking panics during the Great Depression were, thus, special events. Those panics occurred without the private deposit insurance supplied by private bank clearinghouses or the deposit insurance supplied publicly afterwards.

7.2. BANKING PANICS: DESCRIPTION AND THEORIES

A bank panic occurs when depositors demand such a large-scale transformation of deposits into currency that, at the contracted for exchange rate (of a currency dollar for a deposit dollar), the banking system can only respond by suspending convertibility of deposits into currency, issuing clearinghouse loan certificates, or both.[1] Table 7.1 lists the recessions and panics during the National Banking Era, the declines in output as measured by pig iron production, and the increases in the currency-deposit ratio. Also shown are the losses to depositors and the numbers of banks failing. Notice that the banking panics tended to occur just after business cycle peaks. Also, losses on deposits and the number of failures seem small considering that the panics were generalized events which literally involved all banks and depositors.

Two fundamentally different types of theories have been advanced to explain banking panics. The first type of theory, in its traditional form (e.g., Noyes (1909), Gibbons (1968), Kindleberger (1978)), views panics as random manifestations of "mob psychology" or "mass hysteria" rooted in individual and collective psyches. The modern version of the theory that panics are random events is articulated by Diamond and Dybvig (1983), and Waldo (1985). In these models depositors' expectations about the value of deposits are linked to extraneous variables because of an exogenously imposed first-come-first-served rule for bank repurchases of their deposits, in which case the return a depositor receives depends on his place in line at the bank. If the face value of the deposits is larger than the liquidation value of the bank's assets, and there is such a first-come-first served rule, then there exist panic equilibria in which the banking system collapses in panic. Hence, in the Diamond and Dybvig model,

1. Of the seven panics during the National Banking Era five involved suspension of convertibility (1873, 1890, 1893, 1907, 1914) and six involved the issuance of clearinghouse loan certificates (1873, 1884, 1890, 1893, 1907, 1914). During the Panic of 1895 issuance of the loan certificates was authorized, but none were actually issued. Clearinghouse loan certificates are explained in Gorton (1985B) and Gorton and Mullineaux (1986). This definition is much more precise than others which include the nebulous idea of "periods of financial stringency." See, for example, Sprague (1915) and Kemmerer (1910). To be clear, a *bank run* refers to a situation in which depositors at a *single* bank seek to exchange their deposits for currency. A *banking panic* refers to the situation in which depositors at *all* banks want to withdraw currency.

for example, "... anything that causes [depositors] to anticipate a run will lead to a run." Possible causes include "a bad earnings report, a commonly observed run at some other bank, a negative government forecast, or even sunspots" (p. 410). I will subsequently refer to these alleged panic-causing events as "sunspots."

The second type of theory advanced to explain panics argues that panics are systematically related to the occurrence of other events which change perceptions of risk. If there is an information asymmetry between banks and depositors because bank assets and liabilities are nontraded, for example, then depositors might not be able to accurately assess the risk of individual bank's liabilities. They may be forced to use aggregate information. There are three versions of this theory, differentiated by what the relevant aggregate information is taken to be. These theories are: (i) panics are caused by extreme seasonal fluctuations (referred to here as "the Seasonal Hypothesis"); (ii) panics are caused by the (unexpected) failure of a large (typically financial) corporation (referred to as "the Failure Hypothesis"); (iii) panics are caused by major recessions (referred to as "the Recession Hypothesis"). As discussed below, these three hypotheses are not mutually exclusive.

The view that panics are manifestations of seasonal "crises" or seasonal "stringency" was first put forth by Jevons (1884) with reference to England, and later, by Andrew (1906) and Kemmerer (1910) for the United States. Kemmerer identified the seasons when the money market was most "strained" as the periods of the "spring revival" (March, April, May), and the crop-moving period of the fall (September, October, November). He points out that, of the six panics prior to 1910 (the date his work was published), four occurred in the fall and two occurred in the spring. In each case, Kemmerer cites high interest rates, depressed stock prices, and the failure of specific firms as the seasonal effects precipitating panics. He concluded that "the evidence . . . points to a tendency for the panics to occur during the seasons normally characterized by a stringent money market" (p. 232). Andrew (1906) expresses a similar view, and Miron (1985) presents a modern articulation of this traditional view.

The Failure Hypothesis cites the unexpected failure of a large, typically financial, institution as the immediate cause of panics.[2] The argument of the Failure Hypothesis appears to be that because of an information externality such failures created distrust in the future solvency of all banks, leading to withdrawals as depositors sought to avoid expected capital losses on deposits. Since there are many examples of failures of large firms which did not result in panics, a failure

2. The failures cited by contemporary observers of panics and subsequent researchers are as follows: 1873: Jay Cooke and Co.; 1884: Grand and Ward; 1890: Decker, Howell and Co.; 1893: The National Cordage Co.; 1907: The Knickerbocker Trust Co.; 1914: the closing of the stock exchange. Details can be found in the *Commercial and Financial Chronicle* and in many secondary sources.

Table 7-1. NATIONAL BANKING ERA PANICS

NBER Cycle Peak-Trough	Panic Date	%Δ $\left(\frac{C}{D}\right)^*$	% ? Pig Iron †	Loss Per Deposit $ †	% and # Nat'l Bank Failures†
Oct. 1873–Mar. 1879	Sept. 1873	14.53	−51.0	0.021	2.8(56)
Mar. 1882–May 1885	Jun. 1884	8.8	−14.0	0.008	0.9(19)
Mar. 1887–Apr. 1888	No Panic	3.0	−9.0	0.005	0.4(12)
Jul. 1890–May 1891	Nov. 1890	9.0	−34.0	0.001	0.4(14)
Jan. 1893–Jun.1894	May 1893	16.0	−29.0	0.017	1.9(74)
Dec. 1895–Jun. 1897	Oct. 1896	14.3	−4.0	0.012	1.6(60)
Jun. 1899–Dec. 1900	No Panic	2.78	−6.7	0.001	0.3(12)
Sep. 1902–Aug. 1904	No Panic	−4.13	−8.7	0.001	0.6(28)
May 1907–Jun. 1908	Oct. 1907	11.45	−46.5	0.001	0.3(20)
Jan. 1910–Jan. 1912	No Panic	−2.64	−21.7	0.0002	0.1(10)
Jan. 1913–Dec. 1914	Aug. 1914	10.39	−47.1	0.001	0.4(28)

*Percentage change of ratio at panic date to previous year's average.
†Measured from peak to trough.
Data sources provided in Appendix.

per se cannot be the cause of a panic. Writers arguing the Failure Hypothesis generally point to the economic context in which the failure occurs. In general, the economic context of the failure cited is a recession.

The Recession Hypothesis emphasizes that panics occurred as features of severe recessions, presumably because depositors expected large numbers of banks to fail during recessions. During the National Banking Era every major business cycle downturn was accompanied by a banking panic. During this period seven of the eleven cycles (in the NBER chronology) contain panics (see table 7.1). Writers articulating the Recession Hypothesis include Mitchell (1941) and Fels (1959). Mitchell, for example, argues that, "when prosperity merges into crisis . . . heavy failures are likely to occur, and no one can tell what enterprises will be crippled by them. The one certainty is that the banks holding the paper of bankrupt firms will suffer delay and perhaps a serious loss on collection" (p. 74). Like Mitchell, Fels (1959, p. 224) sees panics as "primarily endogenous" parts of the business cycle. Gorton (1985A, 1987A) presents a model of the Recession Hypothesis.

The central common element of all these theories of banking panics is the hypothesized existence of an information asymmetry between banks and depositors which creates the possibility of (information) externalities which change perceptions of the risk of bank deposits, sometimes to the point of panic (e.g., Diamond and Dybvig (1983), Gorton (1987A)). Different explanations of banking panics differ on what variables change perceived risk, but agree that because of the information asymmetry the banking system cannot respond by adjusting the rate of return on deposits. Instead, if there is a panic, the banking system responds to the change in perceived risk by suspending convertibility of deposits into currency rather than adjusting the rate of return. (See Gorton (1985A).) This is because, due to the information asymmetry and consequent externalities, either the change in perceived risk is unrelated to "fundamentals" or it is not possible to credibly raise the rate of return.

7.3. THE DEPOSIT-CURRENCY RATIO

The view that panics are random events places no testable restrictions on the data. Consequently, the basic strategy of analysis followed here is to empirically examine a description of depositor behavior and test whether this description explains depositor behavior at panic dates. In this section the model to be examined is discussed and the hypotheses to be tested are explained. As Miron (1986) points out, data limitations severely constrain the sophistication of models of panics which can be feasibly tested. This section first presents some theoretical motivation for a subsequent, basically *ad hoc*, model which will be estimated.

Consider the behavior of a representative consumer who lives in a Baumol–Tobin economy where consumption goods must be purchased with currency and where "trips" to the bank are costly. Let the number of trips chosen be m_t; let X_t be real consumption, and let p_t be the price level. Under the usual Baumol–Tobin assumptions, currency (C) and deposit holdings (D) during period t are defined as follows:[3]

$$C_t \equiv X_t\,(1/m_t)\,p_t \qquad D_t \equiv X_t\,(1-1/m_t)\,p_t;$$
$$\overline{C}_t = \left(\frac{1}{2}\right) C_t; \qquad \overline{D}_t = \left(\frac{1}{2}\right) D_t;$$

These definitions follow Baumol–Tobin in imposing a binding cash-in-advance constraint on the consumer. For simplicity deposits are the only way of saving.

The representative consumer finances current consumption (X_t) and "trips" (m_t) with last periods' savings and income:

$$\underset{m_t}{\text{MAX}} : E_t \left\{ \sum_{i=t}^{\infty} \beta^{i-t} U\,(X_i)\,|I_t \right\} \tag{I}$$

subject to:

$$X_t + \alpha m_t \leq (1 - r_{dt-1} - \pi_{t-1}) \frac{\overline{D}_{t-1}}{p_t} + Y_{t-1}$$

where:

α is the real cost of a trip;

r_{dt-1} is the real rate of return promised *ex ante* by banks on an average balance deposit dollar held during $t-1$;

π_{t-1} is the real capital loss on an average balance deposit dollar;

Y_{t-1} is real income earned during $t-1$;

β is the subjective rate of time preference;

I_t is the information set available at time *t*.

The budget constraint requires current consumption and current "trip" costs to be financed by income earned (Y_{t-1}) and the return on savings, which is the realized return on the average deposits held last period $\left(\overline{D}_{t-1}\right)$. Since the cash-in-advance constraint is assumed binding, choice of m_t determines current consumption and, simultaneously, choice of savings (through choice of $\overline{D}_t$).

3. The usual assumptions are that "trips" are evenly spaced and that deposits are only drawn down when currency balances are exhausted. See Tobin (1956). Notice, also, that it is without loss of generality that the possibility of writing checks, i.e., using deposits as a medium of exchange, is not allowed. This could be included without changing the basic equation.

The first order condition for problem (I) is:

$$\alpha U'_{X_t} = E_t \left\{ \beta U'_{X_{t+1}} (1 + r_{dt} - \pi_t) \left(\frac{1}{2} \right) (X_t) (1/m_t)^2 \, | I_t \right\} \qquad (7.1)$$

which is a stochastic Euler equation. Similar equations have been extensively studied, e.g., Lucas (1978). In this case, solving (7.1) for m_t and using the above definitions, the relation is a money demand function.

Let utility exhibit constant relative risk aversion where A is the coefficient of relative risk aversion. Then, solving (7.1) for m_t and using the solution in the above definitions, the deposit-currency ratio is obtained:

$$\left[\frac{\overline{D}_t}{\overline{C}_t} + 1 \right]^2 = E_t \left\{ \beta \left(\frac{X_{t+1}}{X_t} \right)^{-A} \frac{\left(\frac{1}{2} \right) X_t}{\alpha} (1 + r_{dt} - \pi_t) \, | I_t \right\} \qquad (7.2)$$

Or, alternatively, letting $S_t \equiv \beta (X_{t+1}/X_t)^{-A} \left(\frac{1}{2} \right) (1/\alpha) X_t$, the deposit-currency ratio can be expressed as:

$$\begin{aligned} \left[\frac{\overline{D}_t}{\overline{C}_t} + 1 \right]^2 &= E_t \{ S_t \, | I_t \} \, E_t \{ (1 + r_{dt} - \pi_t) \, | I_t \} \\ &\quad + \mathrm{COV} \left(S_t ; \, (1 + r_{dt} - \pi_t) \, | I_t \right) \end{aligned} \qquad (7.3)$$

Equation (7.3) is the basic description of the deposit-currency ratio to be studied. In (7.3) the deposit-currency ratio is a function of expectations about the rate of return on demand deposits, the intertemporal terms-of-trade, S_t, and the covariance between the two. An important feature of (7.3) is the specification that the covariance is not time invariant. It depends on the depositor's information. The task is to determine the information on which the expected rate of return and covariance are conditioned, and what, if any, information variables can be identified as causing changes in either the expected rate of return or covariance, such that the deposit-currency ratio declines to the extent of panic. The model does not explain panics, but offers a simple way of embedding the previous discussed models in a single, testable, framework.

7.3.1. The Empirical Model

Equation (7.3) contains a number of unobservable parameters. In particular, A, α, and β are not observable. There are, also, severe data problems. For the nineteenth century, there are no data on the promised rate of return (r_{dt}), the capital loss (π_t), consumption, or demand deposits. Data on currency are incomplete. An additional problem is that the data are not evenly spaced, as explained below. In principle, equation (7.2) could be estimated using the method of moments

(Hansen and Singleton (1982)), ignoring the data problems by using proxies and constructed data. The fact that no consumption data is available, and that the banking data is unevenly spaced seem particularly troublesome. The proxy for consumption data is pig iron production, discussed below. Nothing can be done about the uneven data spacing, except to insure that, as far as possible, data are at the same, if unevenly spaced, dates. Given these problems the moment condition, implied by (7.2), is likely to be misspecified.

These considerations lead to the empirical strategy adopted here. In particular, most of the analysis is conducted using nonparametric methods, after projecting various measures of the covariance and rate of return on possible information variables to get expected values. However, some *ad hoc* versions of equation (7.3) will also be analyzed. Given the use of pig iron production as a proxy for consumption, the resulting equation is best viewed as a reduced form. It is worth stressing, in defense of this approach, that since there are many ways of constructing the different variables required, the reasons why different combinations of constructed variables produce robust results are laid bare.

A basic version of the *ad hoc* model to be estimated is:

$$\left[\frac{\overline{D}_t}{\overline{C}_t}+1\right]^2=\alpha_0+\alpha_1 t+\alpha_2 t^2+\alpha_3(1+r_{dt}-\pi_1^2)+\alpha_4\mathrm{COV}_t^e+\mu t \qquad (7.4)$$

$$\begin{aligned}\alpha_3 &\equiv \mathrm{EXP}\left[\beta_1 \ln\left(X_{t+1}/X_t\right)+\beta_2 \ln X_t\right]\\ \pi_t &= Z_t\gamma+\varepsilon_t \text{ if } Z_t\gamma+\varepsilon_t>0\\ &= 0 \text{ if } Z_t\gamma+\varepsilon_t\le 0\end{aligned} \qquad (7.5)$$

$$\begin{aligned}\mathrm{COV}_t \equiv (X_{t+1}-X_t)\pi_t &= W_t\delta+u_t \text{ if } Z_t\gamma+\varepsilon_t>0\\ &= 0 \text{ if } Z_t\gamma+\varepsilon_t\le 0\end{aligned} \qquad (7.6)$$

The total expected rate of return on demand deposits consists of two components, the "promised" component (r_{dt}) and the expected capital loss component $\left(\pi_t^e\right)$. The promised component is known at time t because it is contractually agreed upon by banks and depositors *ex ante*. Since demand deposits never earn capital gains, the capital loss component (π_t), realized at the end of period t, is constrained, in (7.5), to be positive or zero. Hence, equation (7.5) will be estimated using Tobit methods. The expected capital loss, estimated from (7.5), then enters equation (7.4). In equation (7.5) $\mathbf{Z}_t$ is a matrix of predictors of the capital loss.

Perceived risk is taken as the estimated value of COV_t from equation (7.6) and entered into equation (7.4).[4] In an abuse of terminology, the representative

4. Joint estimation of the model, (7.4)–(7.6), is not econometrically feasible because of the truncation of (7.5), and the affect of that truncation on (7.6). Gorton (1987B) reports on some joint

depositor's conditional forecast of how consumption and losses on deposits will covary is indicated as "expected covariance," COV_t^e. As shown, the asymmetry of the contract, namely that $\pi_t \geq 0$, must be taken into consideration when estimating (7.6). Note that equation (7.6) forecasts only part of the covariance term, the expected product of the change in consumption and the capital loss. The term $1 + r_{dt}$ is not included, and the cross product of the means is not present. The cross product of the means of the change in consumption and the capital loss term can be computed separately. This requires predicting $X_{t+1.}$. Computing the covariance is difficult because of the data problems mentioned above. In fact, none of the relevant data are available. The basic strategy adopted in this paper is to compute a weighted capital loss, as in (7.6), as a proxy. While Gorton (1987B) contains many other results, results reported here all use variations of equation (7.6). In equation (7.6), $\mathbf{W}_t$ is a matrix of predictors, possibly different from $\mathbf{Z}_t$.

Note that while next period's consumption, $X_{t+1,}$ appears in (7.4), the model does not contain an equation predicting next period's consumption.[5] As discussed below, pig iron production is used as a proxy for consumption. Pig iron production can be directly substituted for consumption. Or it may be reasonable to think of current pig iron production as the best estimate of next period's consumption. Subsequently, both possibilities are investigated.

Finally, equation (7.4) contains time trends because data limitations require that deposits be restricted to nationally chartered bank deposits, a declining fraction of total deposits (which include state banks).[6]

7.3.2. Data Considerations

The covariance term is constructed as the weighted loss on deposits, where the weights are the difference in consumption. Pig iron production is used in place of consumption. In what follows a great deal of attention is focused on the covariance term so it is worth briefly discussing each of its components in detail.

Consumption data for the nineteenth century are not available at observation intervals of less than a year. The available annual data are constructed. One possibility would have been to distribute the annual series across months using related series. This would mean using related monthly series to distribute a constructed

estimates of (7.4) and (7.6) when the truncation of (7.5)'s affects on (7.6) are ignored. No results, subsequently reported, seem to turn on this issue.

5. Subsequent reported results were not changed when next period's consumption was predicted with an ARIMA model.

6. The time trend squared is included because the dependent variable is squared.

series. In fact, one of the few monthly series available is pig iron production. Consequently, the second possibility of using pig iron in place of consumption was chosen.

How reasonable a proxy is pig iron production? Berry (1978) presents constructed annual personal consumption expenditures in constant dollars (see Berry's Table 7B). Gallman has also constructed annual estimates of the value of goods flowing to consumers (in constant 1860 prices). Gallman's estimates are unpublished, but are described in Gallman (1966). The correlation between Berry's annual consumption series and the annual average pig iron production, i.e., the average of monthly values, is 0.9270. The correlation between Gallman's annual consumption series and the average annual pig iron production is 0.8877. If, instead of averaging monthly pig iron values to obtain an average annual value, the last value is used, then the correlations are 0.8600 and 0.8260 between pig iron and the Berry and Gallman series, respectively. Thus, pig iron production is a very good proxy for real consumption.[7]

There are no data on the capital losses on deposits. The proxy for actual capital losses during the pre-1914 period is the "loss on assets compounded or sold under order of court" for national banks placed in the hands of receivers (see Appendix). In other words, when a bank failed, court appointed receivers would liquidate the bank over a period of years (sometimes ten or so years). Each year in which an asset was sold at some amount below book value, a loss was recorded. This stream of losses was assigned to the date the bank was closed as the capital loss of deposits.

During a banking panic banks suspend convertibility of deposits into currency. Banks' liabilities, however, continued to circulate in the form of loan certificates and certified checks. (See Gorton (1985B) and Gorton and Mullineaux (1986).) During the period of suspension these bank liabilities exchanged at small discounts against government currency. These discounts represent losses to depositors during suspension periods. Below inclusion of such losses does not change any results.

7.3.3. Hypothesis Testing Using the Model

The model, (7.4)–(7.6), will be used to test three types of hypotheses. First, are banking panics systematic events? The basic claim that panics are systematic events requires testing the hypothesis that the characterization of the

7. Data on pig iron production apparently doesn't exist beyond the series reported in Macaulay (1938). His last observation is dated January 1936. *Historical Statistics of the United States*, however, reports a series on pig iron shipments. The correlation between pig iron shipments and total real consumption over 1929–1970 is 0.7360.

deposit-currency ratio estimated using all nonpanic observations (significantly) holds at the panic dates. If panics are random events caused by extraneous events such as sunspots, then the behavior of the deposit-currency ratio at panic dates should *not* be described by relations which hold at other dates. *From this point of view, economic theories of causal relations are not at issue.* Rather, the question is whether a set of correlations which significantly hold at nonpanic dates also hold at panic dates. If the correlations hold at panic dates, panics will be described as *systematic* events.

A stronger type of claim concerns hypothesized economic behavior, the second type of claim to be examined. In particular, are banking panics predictable? The above model is one of risk averse depositors who seek to optimally smooth consumption intertemporally. The model hypothesizes that losses on deposits which come during periods when depositors want to dissave will be given a lot of weight in utility terms. On the other hand, losses on deposits which occur during periods of rising consumption will be given little weight in utility terms. With respect to panics, if depositors expect a coincidence of declining consumption and high capital losses on deposits, then they will seek to withdraw deposits in advance of those periods. They do this in order to avoid the capital loss which they expect to occur during the period in which they expect to dissave. In other words, panics should not only be systematic, but should be *associated with movements in perceived risk predictable on the basis of prior information.* This hypothesis then requires that the predictors of COV_t not include contemporaneous information (unlike the first claim, above). In this case, panics will be said to be *predictable.*

The third, and final, type of claim concerns what is contained in the information set upon which expectations are conditioned. What type of news causes panics? If panics are systematic, and perhaps predictable, then which of the variables predicting π_t and COV_t are important at all points in time and, *if important at all points in time, which are important at panic dates*? In other words, conditional on panics being, at least, systematic, which of the predictors of the capital loss, π_t, and risk, COV_t, are important at panic dates. *Notice that this excludes predictors which are important at panic dates, but not at other dates.* This restriction, then, tests for "sunspots," if "sunspots" are events which do not occur at all dates. Below, however, the first two claims are re-examined by checking whether predictors found to be unimportant as predictors of COV are important at panic dates.

The three hypotheses that panics are predictable are given empirical form by including in the Z_t and W_t matrices of (7.5) and (7.6) variables (and lags) capturing seasonal effects, failures, and recessions. These variables are taken to be exogenous, and, in fact, are exogenous by Granger-causality tests (see Gorton (1987B)). The Seasonal Hypothesis is represented by the rate of interest on commercial paper (from Macaulay (1938)). Short-term interest rates had

strong seasonals during the pre-Fed period (e.g., Kemmerer (1910), Sargent (1971), Shiller (1980)). The inclusion of short-term interest rates is intended to capture the notion of "seasonal stringency" or "seasonal crisis." The Recession Hypothesis is represented by a leading economic indicator, liabilities of failed nonfinancial businesses.[8]

The Failure Hypothesis emphasizes the unanticipated failure of large, usually financial, institutions. This notion is the hardest to quantify. The Failure Hypothesis is represented by unanticipated capital losses on deposits, i.e., the residuals from the Tobit estimation of capital losses (see Gorton (1987B)). This measure seems close to what the failure hypothesis maintains, but it limits attention to national banks and ignores completely the idea that the failure of specific institutions is what counts.

The variables chosen to capture the content of each hypothesis are not pure representations. All three hypotheses, for example, involve business failures, and short-term interest rates reflect more than seasonals. To some extent these effects can be disentangled. Failed nonfinancial business liabilities and short-term interest rates can be deseasonalized. It is, therefore, possible to test whether failed business liabilities have an impact on the risk measure independent of seasonal movements. Similarly, it is possible to test for effects of interest rates independent of seasonals.

7.4. ANALYSIS OF THE NATIONAL BANKING ERA

The National Banking Era (1865–1914) is examined first because this period preceded the existence of the Federal Reserve System and the Federal Deposit Insurance Corporation, two institutions which may be expected to affect depositor behavior. During the National Banking Era national (though not state) banks were required to report a variety of information to the Comptroller of the Currency five times a year. The Comptroller *Reports* provide most of the data to test the hypotheses of the previous section.

An important drawback to using the Comptroller *Reports* is that information was recorded five times a year (at "call dates"). These reporting dates were not the same every year, but fell in different months. The observations, then, are not evenly spaced.[9] *In what follows the data are treated as if they were*

8. The liabilities of failed businesses led peaks by one cycle phase, and led troughs by two cycle phases (Burns and Mitchell (1946)). Neftci (1979) has shown how the predictive ability of leading indicators can be evaluated by applying a test for Granger causality. By such a test the liabilities of failed businesses does not lead pig iron production, but does lead the risk measure.

9. This was precisely the intention of the Comptroller, who while monitoring banks attempted to keep the call dates from being predictable.

evenly spaced. Also, the information is limited to national banks. All data are described in the appendix. The fact that virtually every series is constructed or proxied has some potentially important implications discussed in particulars later.[10]

The first step in estimating the model of the currency-deposit ratio is estimation of the capital loss on deposits, equation (7.5), the fitted value of which enters equation (7.6). The expected capital loss series is the predicted value from an equation estimated using Tobit analysis due to the truncated distribution of π_t. The equation used in what follows contains a constant term, two lags of the capital loss, the contemporaneous and nine lags of the liabilities of failed businesses, and the contemporaneous and four lags of both pig iron production and the interest rate on commercial paper. The results are not sensitive to specification of this equation and details may be found in Gorton (1987B). In fact, subsequent results are not changed significantly, if, instead of predicted values of the capital loss, actual future capital losses are used. *The reason is that panics are not associated with spikes in the capital loss series. There are many dates at which capital losses are much higher*! There is, thus, prior evidence that the timing of the capital losses with respect to changes in consumption, and not just the level of losses, is important.

7.4.1. Estimates of Perceived Risk

The results of estimating equation (7.6) are all contained in Gorton (1987B). Here those results are summarized. Subsequently, predicted values of COV_t will be used so the importance of equation (7.6) lies in what variables are important predictors of perceived risk (COV_t). In this regard, the results are basically robust to how COV_t is defined, and to whether data are deseasonalized or not.[11]

10. The fact that virtually every data series is constructed, proxied, or interpolated raises a large number of issues and makes the possible combinations of estimates very large. Many of the issues are discussed in Gorton (1987B) which is a large companion appendix to this text. In the text here only the sensitive issues are discussed.

11. Recall that since pig iron is being used as a proxy for consumption, as discussed in the main text, there is the question of the appropriate empirical definition of COV_t . Recall that throughout we are restricting attention to definitions in which the cross product of the means component of COV is ignored. Possible definitions are: (1) $\mathrm{COV}_t \equiv (X_{t+1} - X_t)\left(r_{dt} - \pi_t\right)$; (2) $\mathrm{COV}_t \equiv (X_{t+1} - X_t)\pi_t$; (3) $\mathrm{COV}_t \equiv (X_t - X_{t-1})\left(r_{dt} - \pi_t\right)$; (4) $\mathrm{COV}_t \equiv (X_t - X_{t-1})\pi_t$. The main text argued that the last definition is the appropriate definition. The results discussed in the text are basically robust to which definition is used, though the R^2 in the case of the third definition is more than twice the other cases, whether data are deseasonalized or not, and whether contemporaneous predictors are included or not. The high R^2 does not occur in the case of definition (2), though definitions (1) and (2) give similar results. The likely reason is the way r_{dt} was constructed. See Gorton (1987B) for the full set of results.

The best fits are achieved with ten lags of COV, nine lags of the liabilities of failed nonfinancial businesses, and four lags of the commercial paper rate.[12] The *R*-squared's are all in the range of 0.30. (Estimated coefficients, unimportant for purposes here, can be found in Gorton (1987B).[13])

In all cases, the liabilities of failed businesses variables, deseasonalized or not, are always jointly significant. When short-term interest rates are added to the equation, the liabilities variables, deseasonalized or not, remain jointly significant. Seasonality, as captured by the interest rate variables are always jointly significant. But, notably, when the interest rate on commercial paper is deseasonalized, the interest rates are not jointly significant!

Unanticipated capital losses (representing the Failure Hypothesis) do not appear in any of the final equations used because this variable and lagged values were never jointly significant. There is the possibility that the failure of a single institution occurring in conjunction with business failures is what is important, but attempts to separate these effects did not improve the predictive power of the equation.[14]

12. When tests for whether panics are systematic events, as defined in the main text, contemporaneous values of the liabilities of failed businesses and the commercial rate are also included. Contemporaneous values are excluded when analyzing whether panics are predictable on the basis of prior information.

13. A typical example is as follows:

$$
\begin{aligned}
\mathrm{COV}_t = {} & 0.004 + 0.049\mathrm{COV}_{t-1} - 0.185\mathrm{COV}_{t-1} + 0.006COV_{t-3} \\
& (0.002)\ (0.074) \qquad (0.074) \qquad (0.073) \\
& + 0.074\mathrm{COV}_{t-4} - 0.102\mathrm{COV}_{t-5} + 0.089\mathrm{COV}_{t-6} + 0.001\mathrm{COV}_{t-7} \\
& (0.070) \qquad (0.070) \qquad (0.070) \qquad (0.070) \\
& + 0.020\mathrm{COV}_{t-8} + 0.013\mathrm{COV}_{t-9} - 0.128\mathrm{COV}_{t-10} - 1.44\mathrm{BLIA}_t \\
& (0.070) \qquad (0.068) \qquad (0.067) \qquad (1.56) \\
& - 4.34\mathrm{BLIA}_{t-1} + 1.09\mathrm{BLIA}_{t-2} - 7.33\mathrm{BLIA}_{t-3} - 1.1\mathrm{BLIA}_{t-4} \\
& (1.7) \qquad (1.8) \qquad (1.81) \qquad (1.79) \\
& + 2.9\mathrm{BLIA}_{t-5} - 2.2\mathrm{BLIA}_{t-6} + 3.46\mathrm{BLIA}_{t-7} - 2.77\mathrm{BLIA}_{t-8} \\
& (1.77) \qquad (1.77) \qquad (1.77) \qquad (1.76) \\
& + 4.81\mathrm{BLIA}_{t-9} - 0.011\mathrm{COMP}_t - 0.105\mathrm{COMP}_{t-1} \\
& (1.67) \qquad (0.025) \qquad (0.027) \\
& + 0.091\mathrm{COMP}_{t-2} - 0.057\mathrm{COMP}_{t-3} + 0.041\mathrm{COMP}_{t-4} \\
& (0.028) \qquad (0.028) \qquad (0.027)
\end{aligned}
$$

$R^2 = 0.28$; SSE $= 0.0026$; $F = 2.89$; df $= 186$.
BLIA = Liabilities of failed businesses; COMP ≡ interest rate on commercial paper. This example uses nondeseasonalized data.

14. The contemporaneous liabilities of failed businesses, proxying for both effects, would be mismeasured. OLS estimates, columns (1), (2), and (4), would then be biased and inconsistent. Columns (3) and (5) address this potential problem by using instrumental variables. The instruments were the current value and four lags of loans and discounts at national banks and the current value and four lags of Frickey's Index of Production for Transportation and Communication. (See Gorton (1987B) for details.) Inspection of table 7.2 does not reveal any important differences

It is perhaps important to point out that the information used to predict COV and capital losses on deposits separately was available to agents living during the National Banking Era. The liabilities of failed businesses were published as were interest rate data. In addition, the telegraph, invented in the 1840s, had spread nationwide by the National Banking Era.

7.4.2. Test Results for the Deposit-Currency Ratio Equation

The main results of interest are estimates of the nonlinear deposit-currency ratio equation, (7.4), using predicted perceived risk measures, and expected capital loss measures from the Tobit procedure. Table 7.2 presents a sample of the results. Table 7.2 considers a variety of different COV predictions. In table 7.2, rows (1), (2) and (5) use nondeseasonalized data to predict COV; the remaining rows use deseasonalized data. Rows (1)–(4) use contemporaneous variables, as well as lags, to predict COV; rows (5) and (6) only uses lagged variables.[15]

Consider the first hypothesis to be examined: that panics are systematic events. Table 7.2 addresses this issue by including a dummy variable for the panic dates. If the estimated model cannot explain panics then the dummy variable should be significant. But, the dummy is not significant.[16] The implication is that nothing is happening at panic dates which is not being explained by the model. This conclusion is very strong. It does not depend on the definition of COV, on whether data are deseasonalized, on whether contemporaneous predictors of COV are used, or on the functional specification of the deposit-currency equation. (See Gorton (1987B).)

The evidence is also strong that panics are predictable on the basis of prior information. In table 7.2, the perceived risk variable is significant in all equations. In particular, it is significant when the contemporaneous predictors of COV are omitted as in rows (5) and (6). This means that if, on the basis of prior information, COV_t^e is negative, then depositors shift from deposits to currency in

when the instruments are used, but subsequently the Failure Hypothesis is reexamined, and the measures of perceived risk estimated using instruments.

15. Rows (1) and (3) use instruments to predict COV, as discussed above in footnote 14.

16. The dummy variable is set to one at the panic dates and zero otherwise. The panic dates *in the data* are: December 26, 1873; June 20, 1884; December 19, 1890; July 12, 1893; October 6, 1896; December 3, 1907; September 12, 1914. The results are not sensitive to perturbations of these dates. The dummy variable was not significant in any functional specification attempted. In a log-linear deposit-currency ratio equation, reported on in Gorton (1987B), dummies for the individual panic dates were never significant, individually or as a group. The nonlinear estimation procedure would not converge when individual panic dummies were included.

Table 7-2. DEPOSIT-CURRENCY RATIO TEST RESULTS, 1870–1914

$$\left[\frac{D_t}{C_t}+1\right]^2 = \alpha_0 \text{Dummy} + \alpha_1 + \alpha_2 t + \alpha_3 t^2 + \text{Exp}\left[\alpha_4 \ln\left(X_{t+1}/X_t\right) + \alpha_5 \ln X_t\right]\left(1 + r_{dt} - \pi_t^e\right) + \alpha_6 \text{COV}_t^e$$

	α_0	α_1	α_2	α_3	α_4	α_5	α_6	R^2	σ
1)	0.6349	3.74	−0.0084	0.0002	−0.1971	0.9789	85.99	0.9686	0.4681
	(0.5731)	(0.151)	(0.0029)	(0.00002)	(0.2619)	(0.0304)	(19.89)		
2)	0.5495	3.737	−0.0081	0.0002	−0.1901	0.982	110.34	0.9689	0.4660
	(0.5530)	(0.15)	(0.0029)	(0.00002)	(0.2601)	(0.0303)	(24.26)		
3)	−0.4526	3.689	−0.0079	0.0002	−0.066	0.9881	53.15	0.9668	0.4820
	(0.5044)	(0.1555)	(0.0030)	(0.00002)	(0.2668)	(0.0303)	(21.14)		
4)	−0.5260	3.696	−0.0076	0.0002	−0.0849	0.9882	87.35	0.9673	0.4778
	(0.4918)	(0.1534)	(0.0029)	(0.00002)	(0.2623)	(0.030)	(27.61)		
5)	0.3765	3.729	−0.0080	0.0002	−0.1440	0.9831	101.69	0.9684	0.4690
	(0.5485)	(0.5485)	(0.0029)	(0.00002)	(0.2612)	(0.0303)	(24.42)		
6)	−0.7739	3.712	−0.008	0.0002	−0.0315	0.9844	90.93	0.9671	0.4792
	(0.4888)	(0.1553)	(0.003)	(0.00002)	(0.2661)	(0.0307)	(30.68)		

Standard errors are in parentheses. Dummy = 1 at panic dates and zero otherwise.

order to avoid the capital loss which they expect to occur when consumption is declining.[17] This conclusion is slightly sensitive to the definition of COV_t.[18]

The final caveat concerns the functional form of the deposit-currency ratio equation in the face of the multitude of data assumptions that have been made. Unfortunately, because of the presence of the time trends in the deposit-currency equation, White (1981, 1982) specification tests are inappropriate.[19] The results here are robust to a number of other specifications, however. (See Gorton (1987B).)

Considering the multitude of assumptions about data construction, variable definition, and specification of functional form, and the fact that many of the usual tests cannot be conducted, the robustness of the results is, perhaps, more suspect than usual.[20] Nevertheless the robustness of the results is worth stressing. It seems difficult to argue that there is something special about panics in the sense that the above specification of consumer behavior does not capture behavior during panics. However, the next section re-analyzes the data by concentrating on the panic dates, and avoiding, at least, the specification of the deposit-currency ratio equation. In that sense, the tests in the next section are nonparametric. Such tests also allow for a more precise, and intuitive, sense of what is happening during a banking panic.

17. There is an important data timing problem, discussed subsequently in the main text, which slightly colors these results. The quarterly liabilities of failed businesses observations were assigned to the nearest call date (and the missing value estimated) because of seasonals. In order to avoid mixing up seasons, the resulting series sometimes assigns future values to the current date and sometimes past values. This means that, strictly speaking, including the contemporaneous business liabilities variable as a predictor of COV_t is not inconsistent with the hypothesis that panics are predictable on the basis of prior information.

18. In particular, when r_{dt} is included, the perceived risk measure is not significant. See Gorton (1987B).

19. One possible way to circumvent the problem is to first detrend the data and then test the functional specification. This biases the test in favor of rejection since the White test is now testing the joint hypothesis of correct specification of the detrending function and correct specification of the deposit-currency ratio equation. Gorton (1987B) reports the results of this procedure. In general, the joint hypothesis of correct specification is not accepted.

20. Entering the perceived risk measure into the deposit-currency ratio equation imposes a set of restrictions on the manner in which the predictors of the risk measure are allowed to influence the deposit-currency ratio. If the measure of perceived risk is appropriate, then the imposition of the restrictions should not significantly worsen the fit of the deposit-currency ratio equation. It is well-known that such cross-equation restrictions can be tested (e.g., Barro (1981)). In effect, the test is for whether there is additional information in the predictors of COV_t which affects the deposit-currency ratio through some channel other than perceived risk. Unfortunately, this type of test is inappropriate here because it is not possible to impose the restriction that $\pi_t \geq 0$, i.e., that there are no capital gains to deposits. That is, the truncated value of COV_t, shown in (7.6), cannot be imposed. Gorton (1987B) discusses this issue in greater detail and conducts some experiments concerning its importance. It turns out to be important, so the cross-equation restriction tests are not conducted.

7.5. THE TIMING AND SEVERITY OF PANICS

In this section the actual panic dates are the focus of attention. By focusing on the panic dates it is possible to identify anything "special" which may have occurred. To confirm the above hypotheses, it should be the case that the special event is a large change, a spike, in a variable predicting COV, which in turn, causes a change in the deposit-currency ratio. The special event is the arrival of information which causes depositors to reassess the riskiness of deposits, and to withdraw currency from banks as a consequence. In this section, the channel of causation is analyzed. It is shown that panics did correspond to spikes in the predictors of deposit riskiness, but in a rational way.

The hypotheses that panics are systematic and predictable have testable implications for the timing and severity of panics. With respect to the timing of panics, the hypotheses imply that *at the panic dates* there should be specific, identifiable, movements in the predictors of risk which result in movements in perceived risk and, hence, in the deposit-currency ratio. *Movements in the predictors at panic dates should imply that the perceived risk variable achieves some critical (negative) value at the panic dates. Also, the movements in the risk predictors and in perceived risk should occur at panic dates and not at other dates.* If such movements occurred at other dates, then there should have been panics at those dates.[21]

At the panic dates the magnitudes of the movement of variables can be tested. In effect, the flow of information through the channel of perceived risk at panic dates can be tested. If the information in the predictors of risk is accurate, then the severity of the panic should be related, through the perceived risk measure, to measures of the information content of the predictors. The larger the movement in the predictors, and hence the larger the movements in perceived risk, the larger should be the movements in the deposit-currency ratio.

In addition, if the movements in the predictors are accurate, then the size of these movements, and the associated movements in perceived risk, should be statistically related to the magnitude of downturns in income, rises in capital losses, and the risk measure. The size of the movement in the deposit-currency ratio should be related, through the channel of perceived risk, to the size of income declines and to capital losses. Each of the three hypotheses about what the relevant predictive information is can be examined with respect to the above implications for the predictors of risk.

21. This statement, however, is subject to an important caveat. Following panic dates deposits may be perceived as even riskier, as depositors get more information, but depositors have already withdrawn their deposits, the banking system has suspended convertibility, or depositors have converted their deposits into clearinghouse loan certificates. On clearinghouse loan certificates see Gorton (1985B), Gorton and Mullineaux (1986), and Cannon (1910).

7.5.1. Tests of Timing Relations

At a panic date the perceived risk variable should achieve a critical or threshold value not achieved at other dates. Using five different measures of perceived risk, table 7.3 lists the number of times the perceived risk measure achieved a lower value *before* the panic date (i.e., previous business cycle peak to panic date) and after the panic date (i.e., panic date to subsequent business cycle trough).[22] As a reference, the first column of the table lists the number of data points between the previous peak and the panic date (labelled "before") and between the panic date and the subsequent trough (labelled "after"). The results are quite striking: negative spikes in the perceived risk measures tend to occur at panic dates.

Is there a threshold value of perceived risk which, when reached, results in a panic? The evidence, while sensitive to the perceived risk measure, supports the existence of such a critical value. In the case of the first perceived risk measure, $\mathrm{COV}_t^e(1)$, for example, there are a total of four values lower (i.e., "more" negative) than those occurring at the panic dates, three associated with the Panic of 1884. $\mathrm{COV}_t^e(2)$ also has some problem with the Panic of 1884. The last three perceived risk measures indicate that spikes do, indeed, tend to occur at the panic dates. It is rare for there to be a spike in the perceived risk variable before or after the panic.

What causes the large negative values or spikes in the perceived risk measure at panic dates? Do these spikes correspond to identifiable movements or spikes in the predictor variables? In order to test these implications for the three hypotheses, measures of the information content of the (contemporaneous) predictors of perceived risk are needed. Three measures of the liabilities of failed businesses are used in subsequent tests. The first measure attempts to capture the new information in the liabilities of failed businesses, movements in the variable not predictable on the basis of prior information (its own history). This measure is unanticipated changes in the liabilities of failed business (UNLIA), measured by the residuals from an estimated ARIMA model (see Gorton (1987B)). The second measure is the cyclic component of the liabilities of failed businesses series (CCBUS), measured as the log of the observation minus the mean of the logged series. The third measure, using deseasonalized data, is the observation minus the mean of the series (DECC).

22. The five measures of perceived risk all define $\mathrm{COV}_t \equiv (X_t - X_{t-1})\pi_t$, where X_t is pig iron production at date t, and π_t is the capital loss on deposits at date t . See footnote 11. All the equations use the lags of COV, nine lags of the liabilities of failed businesses, and four lags of the commercial paper rate. In table 7.3, $\mathrm{COV}_t^e(1)$, $\mathrm{COV}_t^e(2)$, and $\mathrm{COV}_t^e(3)$ use nondeseasonalized data, $\mathrm{COV}_t^e(4)$ and $\mathrm{COV}_t^e(5)$ use deseasonalized data. $\mathrm{COV}_t^e(5)$ was estimated jointly with the deposit-currency ratio equation. The estimated equations are provided in Gorton (1987B).

Table 7-3. TIMING OF MEASURES OF PERCEIVED RISK[†]

Panic of	#Data Points		$COV_t^e(1)$		$COV_t^e(2)$		$COV_t^e(3)$		$COV_t^e(4)$		$COV_t^e(5)$	
	Before	After	Before*	After	Before	After	Before	After	Before	After	Before	After
1873	0	26	0	0	0	0	0	0	0	2	0	0
1884	13	4	3	1	6	3	1	1	1	1	1	1
1890	2	2	0	1	1	1	0	0	1	1	0	0
1893	2	4	0	1	0	1	0	1	0	1	0	0
1896	4	3	0	0	0	1	0	0	0	0	0	0
1907	2	2	0	1	0	1	0	0	0	0	0	0
1914	8	2	1	1	2	1	0	0	0	0	0	0

* Number of times the perceived risk measure is lower, i.e., "more negative," than the value at the panic date, previous peak to from panic date (Before), and from panic date to subsequent trough (After).

[†] The five measures, $COV_t^e(1)$, $COV_t^e(2)$, etc., are defined in footnote 22.

The commercial paper rate is examined by looking at deviations from seasonals. In other words, at panic dates the observed rate of interest should be higher than the expected seasonal movement. Such a deviation is intended as a measure of "seasonal stringency." The unanticipated losses on deposits, intended to capture the Failure Hypothesis, are also re-examined.

First, the timing of movements in the liabilities of failed businesses predictor are examined. Table 7.4 lists the largest positive values of unanticipated increases in the liabilities of failed businesses (UNLIA) and the largest positive values of the cyclical component of liabilities of failed businesses (CCBUS for nondeseasonalized data; DECC for deseasonalized data). *In each case there are no positive shocks larger than those listed in the table.* For each measure of the information in the liabilities variable, the values listed are equal to or higher than the lowest value at a panic date.

The results in table 7.4 are striking: panics tend to correspond to the largest values of the liabilities shocks. By the CCBUS measure, every time a shock greater than or equal to 0.8264 occurred after a business cycle peak, there was a panic. Also, the panics correspond to the first large shock following the latest business cycle peak. There are some exceptions. For example, by the UNLIA measure, the shock in November 1887 did not cause a panic, while a smaller one did in June 1884.[23]

The deviation of the commercial paper rate from its seasonals is positive at all the panic dates, but there are *larger* deviations at many other dates. In fact, at 33 nonpanic dates there are positive deviations higher than the lowest positive deviation at a panic date. Nor is there any particular (e.g., business cycle) pattern to the seasonal shocks. This evidence suggests that seasonality in interest rates is not important for panics, though it is important for movements in perceived risk and, hence, the deposit-currency ratio over the whole cycle.

The results for unanticipated losses on deposits are similar to those for seasonal deviations in the commercial paper rate. At three of the panic dates there were no unanticipated losses on deposits. At eight nonpanic dates the unanticipated losses were higher than the *highest* unanticipated loss at a panic date. There are many cases of positive unanticipated losses, with no apparent pattern. By this measure the Failure Hypothesis again seems unimportant. The timing evidence

23. The Panic of 1895–96 was the mildest panic of those discussed and constitutes a borderline case. The New York Clearing House Association authorized the use of loan certificates on December 23, 1895, but no member banks applied for them. In late August 1896 the loan certificate process was again activated in response to panic conditions. (See New York City Clearing House Loan Committee *Minutes*.) The *Commercial and Financial Chronicle* describes September to December 1895 and December 1896 as "panicky periods." Spikes in the liabilities variable in October 1896 would then be accurate since December 1896 is not a data point.

Table 7-4. TIMING OF LIABILITIES OF FAILED BUSINESSES SHOCKS

NBER Chronology Peak Trough	Largest Values of UNLIA	Largest Values of CCBUS	Largest Values of DECC	Panic Date
Oct. 1973–Mar. 1879	(Sep. 1873: 1.1474) Dec. 1873: 1.5028 (Feb. 1874: 1.272)* (Oct. 1878: 0.7587) (Oct. 1883: 0.782)	Dec. 1873: 1.4012 (Feb. 1874: 1.1511) (Mar. 1878: 0.9397)	Dec. 1987: 0.08187 (Feb. 1874: 0.05181)* (Jun. 1878: 0.04086)	Dec. 1873
Mar. 1882–May 1885	Jun. 1884: 1.0535	Jun. 1884: 0.9653	Jun. 1884: 0.07631	Jun. 1884
Mar. 1887–Apr. 1888	Nov. 1887: 1.307	Nov. 1887: 0.8223		No Panic
Jul. 1890–May 1891	Dec. 1890: 1.1249	Dec. 1890: 1.0216	Dec. 1890: 0.03956	Dec. 1890
Jan. 1893–Jun. 1894	Jul. 1893: 1.4340	Dec. 1893: 1.3323	Jul. 1893: 0.11365 (Jul. 1895: 0.03313)	Jul. 1893
Dec. 1895–Jun. 1897	Oct. 1896: 0.8780	Oct. 1896: 0.8264	Jul. 1896: 0.03255 (Jul. 1897: 0.03579)	Oct. 1896
Jun. 1899–Dec. 1900			Jun. 1900: 0.03383	No Panic
Sep. 1902–Aug. 1904				No Panic
May 1907–Jun. 1908	Dec. 1907: 0.8712 (Feb. 1908: 0.8763)*	Dec. 1907: 0.9308	Dec. 1907: 0.03183	Dec. 1907
Jan. 1910–Jan. 1912	Mar. 1910: 1.041	Mar. 1910: 0.8236 (Apr. 1913: 0.8736) (Jan. 1914: 0.8618) (Jun. 1914: 1.1558)	Jun. 1911: 0.03615 (Jun. 1913: 0.0482) (Mar. 1914: 0.0396) (Jun. 1914: 0.0940)	No Panic
Jan. 1913–Dec. 1914	Mar. 1914:0.7545	(Jun. 1914: 1.1558) Sep. 1914: 0.9958 (Dec. 1914: 0.9863) (Mar. 1914: 0.9535)	Sep. 1914: 0.0434 (Dec. 1914: 0.0986)	Sept. 1914

*During suspension.

with respect to the predictors of perceived risk suggests that for panics the liabilities of failed businesses is the important variable. Banks hold claims on firms, and when firms begin to fail in sufficiently large numbers, it signals the onset of a recession and a panic is likely to occur.

Remarkably, the data support the notion of a critical or threshold value of the liabilities of failed businesses variable, and a threshold value of the perceived risk measure, at the panic dates. The seemingly anomalous event of a panic appears to be no more anomalous than recessions.

7.5.2. Severity Tests

While strongly suggestive, the timing of variables discussed above does not constitute a test. However, Spearman's rank correlation coefficient can be used to test the implications of the systematic hypothesis for timing and severity. The rank correlation test is important because it can check that the above hypotheses explain panics when the data are unconstrained by nonpanic relations. The test is conducted by ranking the measures of the information content of the predictors, the perceived risk measures, the currency-deposit ratio, measures of the severity of recessions, and measures of losses on deposits. The Spearman rank correlation coefficient can then be used to test whether the correlations between the movements of these variables *at the specified dates* are significant.

The results are presented in table 7.5. The Spearman rank correlation coefficients are shown for seventeen variables which were ranked at eleven dates.[24] The first three variables are measures of the severity of the eleven recessions during the National Banking Era. Columns (4) and (5) are the percentage changes in the money stock and currency-deposit ratio for the selected dates through the subsequent recession.[25] Columns (9)–(11) are measures of losses on deposits. Columns (14)–(17) are four measures of perceived risk.[26] The notes to the table explain the other variables.

The results in table 7.5 broadly confirm the earlier conclusion that panics are systematic. The nondeseasonalized measures of failed business liabilities

24. Seven of the dates were the panic dates. The remaining four dates correspond to the remaining four business cycles during the National Banking Era. These four dates were selected according to the largest spikes in the measures of the information in the liabilities variable. The dates used were: December 26, 1873; June 20, 1884; October 5, 1887; December 19, 1890; July 12, 1893; October 6, 1896; June 29, 1900; January 22, 1904; December 3, 1907; March 29, 1910; September 12, 1914.

25. Results are unaffected if percentage changes are computed from peak to trough.

26. The four measures of perceived risk correspond to the first four measures described in footnote 22. Gorton (1987B) contains similar results using other measures of perceived risk.

Table 7-5. SPEARMAN RANK CORRELATION COEFFICIENTS

	Eckler (Overall)	Eckler (Pig Iron)	Achinstein (Amplitude)	$\%\Delta M$	$\%\Delta \frac{C}{D}$	UNLIA	CCBUS	DECC	Losses	Total Losses	Post-panic Losses	RES	DECOMP	COV^e (1)	COV^e (2)	COV^e (3)	COV^e (4)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
(1)		0.90	0.818	0.564	0.609	0.888	0.618	0.582	0.818	0.935	0.913	0.782	0.445	0.636	0.591	0.718	0.636
(2)			0.827	0.50	0.755	0.867	0.668	0.527	0.627	0.909	0.873	0.605	0.455	0.70	0.618	0.691	0.60
(3)				0.65	0.632	0.877	0.682	0.609	0.782	0.944	0.914	0.536	0.591	0.709	0.618	0.645	0.536
(4)					0.564	0.862	0.736	0.755	0.691	0.818	0.764	0.482	0.464	0.445	0.336	0.445	0.409
(5)						0.872	0.70	0.273	0.391	0.914	0.879	0.282	0.727	0.836	0.755	0.773	0.682
(6)							0.65	0.491	0.605	0.956	0.962	0.268	0.482	0.591	0.555	0.536	0.436
(7)								0.673	0.545	0.949	0.91	0.345	0.482	0.809	0.673	0.673	0.60
(8)									0.436	0.60	0.509	0.345	0.264	0.518	0.30	0.355	0.345
(9)										0.791	0.782	0.682	0.527	0.627	0.573	0.664	0.727
(10)											0.989	0.473	0.718	0.759	0.736	0.755	0.709
(11)												0.382	0.627	0.636	0.664	0.645	0.645
(12)													0.355	0.491	0.364	0.582	0.391
(13)														0.855	0.891	0.818	0.80
(14)															0.918	0.882	0.827
(15)																0.982	0.873
(16)																	0.855
(17)																	

The rank correlation can range from −1 (completely uncorrelated) to +1 (perfectly correlated). The correlation is significant at the 5% level if the calculated coefficient is higher than 0.553, and at the 1% level if the calculated coefficient is higher than 0.684.

Columns (l)–(3) are measures of the severity of the eleven recessions during the National Banking Era. Column (1) is an overall measure from Eckler (1933). Column (2), also from Eckler (1933), is a measure based only on pig iron production. Column (3) is a measure of the amplitude of each cycle from Achinstein (1961). Columns (4) and (5) are the percentage changes in the money stock and currency-deposit ratio from the selected date through the trough. Columns (6), (7), and (8) are the measures of the liabilities variable, UNLIA, CCBUS, and DECC, respectively. Column (9) is the loss per deposit dollar at the selected date, while column (10) is the loss per deposit dollar from peak through the trough (Total Losses). Column (11) is the loss per deposit dollar from selected date through the trough (Post-panic Losses). Column (12) is the unanticipated loss on deposits measure (RES). Column (13) is the deviation from the seasonal for the rate of return on commercial paper (DECOMP). Columns (14)–(17) are four measures of perceived risk. (See footnote 26.)

(UNLIA, CCBUS) are significantly correlated with the measures of risk which do not use deseasonalized data (COV^e (1), COV^e (2)). The cyclical component, CCBUS, is significantly correlated with all the measures of perceived risk.[27] The deviations of the commercial paper rate from its seasonals (DECOMP) are significantly correlated with all the measures of perceived risk, though not with any measure of the liabilities variable. The unanticipated losses on deposits (RES) are correlated with one measure of perceived risk, $COV^e(3)$.

The business cycle aspect of panics is also revealed again. The percentage change in the currency-deposit ratio is significantly correlated with all the measures of perceived risk. Both the currency-deposit ratio and the perceived risk measures are significantly correlated with the measures of recession and losses on deposits.

The results of this section confirm the earlier conclusion that panics are systematic. The stronger hypothesis that panics are predictable is problematic. Causal inferences would be stronger if it could unambiguously be stated that panics are predictable on the basis of *prior* information, rather than on the basis of contemporaneous information. But there is an important data timing problem. The quarterly liabilities of failed businesses observations were assigned to the nearest call date (and the missing value estimated) because of seasonals. The resulting series then sometimes assigns future values to the current date and sometimes past values.[28] If the contemporaneous value of the liabilities variable is omitted in equations (7.5) and (7.6), estimates of equation (7.4) are basically unaffected, but the analysis *at the panic dates* using only lags of the liabilities variable to predict COV_t results in insignificant correlations.[29] The problem seems

27. Notice, however, that seasonality in the liabilities variable seems important. The nondeseasonalized measure (UNUA, CCBUS) are significantly correlated with the measures of risk, but the deseasonalized liabilities measure (DECC) is *not* significantly correlated with any of the perceived risk measures.

28. Three dates are relevant: the actual date of the panic; the dating of the Comptroller's Reports; the assignment of the quarterly liabilities of failed businesses variable. The call date in the Comptroller's *Reports* immediately *after* the panic date is assigned to the panic in the data (though "immediately after" varies by up to almost three months). At these call dates, corresponding to the panics, the liabilities variable is dated *after*, but in the same month, in four cases, and before, in the immediately proceeding month in two cases. These were the closest assignments. In the case of the Panic of 1873 the liabilities variable was estimated from railroad bond defaults (see Gorton (1987B)). The problem is further complicated by the fact that the liabilities variable is cumulative over the quarter.

29. More accurately, the perceived risk estimates are often zero at several panic dates, so that there is no way to rank them and conduct the tests. In the one case where this is not true, however, the perceived risk measure is significantly correlated with the percentage change in currency-deposit ratio. See Gorton (1987B).

to be that the liabilities observations lagged once are "too far away." In short, the data are not fine enough to adequately draw stronger inferences.

7.6. THE FEDERAL RESERVE SYSTEM, DEPOSIT INSURANCE, AND PANICS

The Federal Reserve System, begun in 1914, and deposit insurance, initiated in 1934, were both introduced primarily to prevent banking panics. This section examines the effects of these two monetary regimes on depositor behavior by estimating the model over these subsequent periods. All data, estimated equations, and test statistics for this section are detailed in Gorton (1987B).

7.6.1. The Period 1873–1934

The introduction of the Federal Reserve System significantly altered depositor behavior. Both the perceived risk equations and the deposit-currency ratio equations exhibit significant structural changes after 1914.[30] A more precise sense of the difference made by the existence of the Federal Reserve System may be obtained by examining the timing of the measures of the information content of the liabilities of failed businesses variable during the period of 1914–1934. Table 7.6 lists the largest liabilities shocks for the peak to trough phase of the business cycles during this period. The table presents two measures. The unanticipated liabilities measure (UNLIA) was estimated over the period 1873–1934 and is, thus, comparable with the earlier period (table 7.4). The cyclical component of the liabilities shock (CCBUS) was computed as the logged value minus the mean of the logged value over the years 1914–1934.

Examining the table, the UNLIA shock in June 1920 was large enough to have precipitated a panic had it come during the National Banking Era, but there was no panic under the Federal Reserve system. The UNLIA shock in December 1929 also did not precipitate panic, though it would have during the National Banking Era. The December 1929 shock coincides with the stock market crash since October 1929 is not a data point. By the other measure, CCBUS, which is not comparable with the earlier period, there is also a spike in December 1929.

Notably, the timing of the UNLIA shocks in June 1920 and December 1929 are the same as the pre-Fed era. Both shocks come just following the business cycle

30. Tests for structural change after the introduction of the Federal Reserve System, and deposit insurance in 1934, were done on the equations predicting COV_t, the deposit-currency ratio equation, and a log-linear deposit-currency ratio equation. The evidence favored the existence of structural change under all data definitions, using the usual Chow tests.

Table 7-6. TIMING RELATIONS DURING THE PERIOD 1914–1934

Peak-Trough	**UNLIA Shock**	**CCBUS Shock**	**Panic**
Aug. 1918–Mar. 1919	Nov. 1918 0.2435	No Positive Shocks	No Panic
Jan. 1920–July 1921	June 1920 1.1341	Mar. 1921 0.7767	No Panic
May 1923–July 1924	Nov. 1923 0.5199	Mar. 1924 1.1473 (Oct. 1923 0.9392)	No Panic
Oct. 1926–Nov. 1927	Apr. 1927 0.2685	Mar. 1927 0.6584	No Panic
Aug. 1929–Mar. 1933	Dec. 1929 0.7687	Dec. 1929 0.7775	Oct. 1930
		Jan. 1931 1.1157	Mar. 1931
		Jan. 1932 1.1392	Jan. 1933
		Feb. 1932 1.0074	
		Mar. 1932 1.1061	
		Apr. 1932 1.1817	
		Jan. 1933 0.9366	

UNLIA was estimated over the period 1873–1934. CCBUS was estimated over 1914–1934.

peaks. Simple tests on processes generating the failure liabilities do *not* reject the null hypothesis of no structural change (see Gorton (1987B)). In other words, the introduction of the Federal Reserve System did not alter the process driving failure liabilities. Depositor behavior changed. In deposit-currency ratio equations over the 1914–1934 sample period, measures of perceived risk are always insignificant though the perceived risk equations perform *best* over this period (see Gorton (1987B)). The panics of the 1930s happened in October 1930, March 1931, and January 1933, *well after* the business cycle peak. So the existence of the Fed did prevent a panic in June 1920, but only altered the timing of the later panic.

7.6.2. The Period 1914–1972

The introduction of deposit insurance again significantly altered depositor behavior. Both the perceived risk equations and the currency-deposit ratio equations exhibit significant structural changes after 1934. Following the introduction of deposit insurance there were several cases of large failed business liabilities shocks, none of which precipitated panics. Like the results for the 1914–1934 period, the perceived risk measure is insignificant in the deposit-currency ratio equation estimated over the 1935–1972 sample period. Over the 1914–1934 period the sign on the perceived risk measure is positive as it is over the pre-Fed period. That is, in response to an expected coincidence of capital losses on deposits with declining consumption, i.e., $COV_t < 0$, depositors

reduced their deposit-currency ratios. However, over the 1935–1972 sample period the sign on the perceived risk measure is consistent with the success of deposit insurance. Expecting to dissave during recessions, when the perceived risk measure is negative, depositors increased their deposit-currency ratios.

7.7. THE 1920S AND 1930S WITHOUT THE FED

What would have happened during the 1920s and 1930s if the Federal Reserve System had not come into existence? This question can be partly answered if it is assumed that depositors would have reacted to the liabilities of failed businesses signal during the 1920s and 1930s in the same way as during the National Banking Era. Recall that tests of the null hypothesis that the process generating the liabilities variable is not stable over the 1873–1934 sample period are rejected. As previously indicated, the UNLIA shock estimated over the period 1873–1934 is appropriate for the counterfactual. According to this UNLIA series (see table 7.6), there would have been a panic in June 1920, and another panic in December 1929. These panics would have followed the timing pattern of the panics during the National Banking Era. The June 1920 spike comes shortly after the business cycle peak of January 1920 (the trough was July 1921). The December 1929 spike follows the August 1929 peak (trough: March 1933).

To construct the counterfactual, two further reduced form equations must be estimated to characterize the effects of depositor responses to changes in perceived risk during panics. Using the observations on the seven panics during the National Banking Era, the percentage of failing banks in the system and the percentage losses on deposits can be predicted using the UNLIA shock. The estimated reduced form relations are:

$$\begin{aligned} \%\mathrm{FAIL}_t &= 0.010023\,\mathrm{UNLIA}_t \\ &\quad (0.0027) \\ R^2 &= 0.6973 \quad \mathrm{DW} = 1.7019 \ \text{d.f.} = 6 \end{aligned} \tag{7.7}$$

$$\begin{aligned} \%\mathrm{LOSS}_t &= 0.062942\,\mathrm{UNLIA}_t \\ &\quad (0.0204) \\ R^2 &= 0.6129 \ \mathrm{DW} = 1.7097 \ \text{d.f.} = 6. \end{aligned} \tag{7.8}$$

Standard errors are in parentheses. The observations on losses and failures are cumulative from the panic date through the trough date, divided by total deposits and total number of national banks, respectively, at the panic date.

Table 7.7 compares the actual percentages of failures and losses, from the panic dates through the troughs, with the values predicted using (7.7) and (7.8). For the actual percentages of banks failing from December 1929 through March

Table 7-7. Actual and Predicted Losses and Failures, 1920s and 1930s

		Failures	
Date	**Predicted % of National Banks Failing**	**Actual % of National Banks Failing**	**Actual % of All Banks Failing**
June 1920	1.137	0.27	0.91*
Dec. 1929	0.77	26.24; 13.36	36.08; 30.76
		Losses	
Date	**Predicted Losses at National Banks (%)**	**Actual Losses at National Banks (%)**	
June 1920	7.14	0.42	
Dec. 1929	4.84	18.407	

* Covers the period Jan. 1921–July 1921 and uses the number of all banks in June 1921 as the base. Data on all banks begin in 1921.
All data are described in Gorton (1987B).

1933, two numbers are listed. The first uses the Federal Reserve System's definition of suspension which is not strictly comparable (see Gorton (1987B)). The second number, in the case of National Banks, uses the number of receiverships closed during 1930–1933. The second number, in the case of all banks, uses the number of banks which did not reopen after the March 1933 banking holiday (2, 132), instead of the Federal Reserve number for suspensions during March 1933 (3,460) (see Gorton (1987B)). Neither of these measures is strictly comparable. The two numbers, however, are the upper and lower limits. The loss measures, however, are comparable.

Table 7.7 shows that *if there had been a panic in June 1920, the percentages of banks failing and losses on deposits would have been higher than those which actually happened.*[31] *However, if there had been a panic in December 1929, failure and loss percentages would have been an order of magnitude lower.* Losses and failures from June 1920 through the trough (July 1921) were lower than predicted perhaps because there was no panic. Between December 1929 and April 1933, there were three panics which came near the trough (October 1930; March 1931; January 1933). Losses and failures, however, were much higher than predicted. Table 7.7 indicates that the magnitudes of the losses and failures during the 1930s cannot be explained by the relations operating prior to the existence of the Federal

31. Moreover, prior to 1920 state bank failure rates were about three times the rates for National banks (Bremer (1935)). This would increase the differences between the actual and predicted values for June 1920.

Reserve System. The existence of the Federal Reserve System altered depositors' perceptions of risk, as indicated by the insignificance of the perceived risk measures in the deposit-currency ratio equations estimated over the 1914–1934 sample period (see Gorton (1987B)).

7.8. CONCLUSION

The results of this study are a set of stylized facts about banking panics, which, while extremely important since their reoccurrence motivated bank regulation, are not well understood. The main stylized fact is that panics are systematic (as previously defined) events linked to the business cycle. Panics turn out not be mysterious events after all. The evidence favors the conclusion that panics were a manifestation of consumption smoothing behavior on the part of cash-in-advance constrained agents. Panics seem to have resulted from changes in perceived risk predictable on the basis of prior information. The recession hypothesis best explains what prior information is used by agents in forming conditional expectations. Banks hold claims on firms and when firms begin to fail, a leading indicator of recession (when banks will fail), depositors reassess the riskiness of deposits.

Depositors panic when the liabilities signal is strong enough. In fact, during the National Banking Era, whenever the information measure of the liabilities of failed businesses reached a "critical" level, so did perceptions of risk and there was a banking panic. In this sense panics were special events. The cyclical behavior of the liabilities variable made panics an integral part of the pre-1914 business cycle.

As with all statistical inference, the above results cannot reject the notion that there exists an unknown variable(s) causing simultaneous increases in the currency-deposit ratio, risk, and the liabilities of failed businesses. *However, we can say that the influence of such unknown factors must happen the same way at panic and nonpanic dates, which is not consistent with sunspot theories of panics.* Sunspot theories argue that there is something special going on at the panic dates which does not occur at other dates, i.e., sunspots, but this is not consistent with the above evidence.

Could the causality be reversed in the above conclusions? Might it not be the case that depositors panic because of sun spots, run the banks, and thereby, cause the banker to call in loans, causing firms to fail? This scenario can be eliminated for three reasons. First, capital losses on demand deposits do not Granger-cause the liabilities of failed businesses, but liabilities of failed businesses do Granger-cause losses on deposits.[32] In other words, the mechanism

32. In regressions with ten lags of each variable, the F statistic for the liabilities variable with capital losses on deposits as the dependent variable was 2.11 (d.f. = (11, 184)), significant at the 5 percent

of causality running from depositors withdrawing currency from "illiquid" banks and causing businesses to fail is not present, at least when all dates are examined. Second, the response of banks to panics was not to liquidate loans, but to issue circulating private money which insured depositors against the failure of individual banks. (See Gorton (1985B, 1987A).) Finally, call loans do not seem to have been sizable enough to be the mechanism, and do not seem to have been loaned to nonfinancial businesses, in general. (See Myers (1931).)

At the panic dates the important shock seems to be the liabilities of failed businesses (with a seasonal component). This result was the basis of the counterfactual about the 1920s and 1930s. After 1914 the private insurance arrangements of commercial bank clearinghouses were replaced by the Federal Reserve System (see Gorton (1985B)). The counterfactual reveals the inadequacies of drawing policy conclusions about private market failures from the experience of the Great Depression. The evidence suggests that the private insurance arrangements of clearinghouses compare favorably to the Federal Reserve System in responding to banking panics.

APPENDIX

Gorton (1987B) contains complete details of data sources and data construction methods, as well as further results. The basic data sources are as follows. Currency in the hands of the public and demand deposit data are from the *Annual Report* of the Secretary of the Treasury, Friedman and Schwartz (1963), *Survey of Current Business* (Supplements), *Banking and Monetary Statistics*, and the *Annual Statistical Digest* of the Federal Reserve System. The liabilities of failed businesses series is from *Financial Review* and from *Survey of Current Business*, for the later period. Pig iron production is from Macaulay (1938). Capital losses on demand deposits are constructed from the Comptroller *Reports* and from FDIC *Annual Reports.* Data on bank suspensions are from the Federal Reserve *Bulletin*, September 1937. Earlier data on the number of national banks failing are from the Comptroller *Reports* of 1925 and 1935.

REFERENCES

Achinstein, Asher (1961), "Economic Fluctuations," Chapter 6 of Seymour Harris, ed., *American Economic History* (New York; McGraw-Hill Book Company).

level. In other words, the liabilities variable Granger-causes losses. The reverse test results in an F statistic of 0.38. Losses on deposits do not cause business failures.

Andrew, A. P. (1906), "The Influence of the Crops on Business in America," *Quarterly Journal of Economics*, XX.

Barro, R. J. (1981), "Unanticipated Money Growth and Economic Activity in the United States," Chapter 5 of *Money, Expectations, and Business Cycles* (New York; Academic Press).

Berry, Thomas Senior (1978), *Revised Annual Estimates of American Gross National Product*, Preliminary Annual Estimates of *Four Major Components of Demand, 1789–1889*, Bostwick Paper No. 3 (The Bostwick Press).

Bremer, C. D. (1935), *American Bank Failures* (New York; Columbia University Press).

Burns, Arthur, and Mitchell, Wesley (1946), *Measuring Business Cycles* (New York; National Bureau of Economic Research).

Cagan, Phillip (1965), *Determinants and Effects of Changes in the Stock of Money, 1875–1960* (New York; National Bureau of Economic Research).

Cannon, J. G. (1910), *Clearing Houses* (National Monetary Commission, S. Doc. 491, 61st Cong. 2nd sess.).

Chow, Gregory, and Lin, An-loh (1971), "Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series," *Review of Economics and Statistics* 53(4) (November).

Diamond, D., and Dybvig, P. (1983), "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, Vol. 91, No. 3 (June).

Eckler, A. R. (1933), "A Measure of the Severity of Depressions, 1873–1932," *Review of Economic Statistics* (May), XV, 75–81.

Fels, Rendigs (1959), *American Business Cycles, 1865–1897* (University of North Carolina Press).

Friedman, M., and Schwartz, A. J. (1963), *A Monetary History of the United States, 1867–1960* (Princeton University Press).

__(1970), *Monetary Statistics of the United States* (New York; National Bureau of Economic Research).

Gallman, Robert E. (1966), "Gross National Product in the United States, 1834–1909," In *Output Employment and Productivity in the United States after 1800*, Studies in Income and Wealth, Volume 30, by the Conference on Research in Income and Wealth, National Bureau of Economic Research (Columbia University Press).

Gibbons, J. S. (1968), *The Banks of New York, Their Dealers, the Clearing House, and the Panic of 1857* (New York; Greenwood Press; reprint of 1859 original).

Gorton, G. (1985A), "Bank Suspension of Convertibility," *Journal of Monetary Economics* 15(2) (March).

__ (1985B), "Clearing Houses and the Origin of Central Banking in the U.S.," *Journal of Economic History*, 45(2) (June).

__ (1987A), "Incomplete Markets and the Endogeneity of Central Banking," Rodney L. White Center for Financial Research Working Paper #16–87, The Wharton School, University of Pennsylvania.

__ (1987B), "Banking Panics and Business Cycles: Data Sources, Data Construction, and Further Results," The Wharton School, University of Pennsylvania, mimeo.

Gorton, Gary, and Donald J. Mullineaux (1987), "The Joint Production of Confidence: Endogenous Regulation and 19th Century Commercial-Bank Clearinghouses," *Journal of Money, Credit, and Banking*, Vol. 19, No. 4 (November).

Hansen, L. P., and K. J. Singleton (1982), "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models," *Econometrica* 50, 1269–86.

Jevons, Stanley W. (1884), *Investigations in Currency and Finance* (London).

Kemmerer, Edwin W. (1910), *Seasonal Variations in the Relative Demand for Money and Capital in the United States* (Washington, D.C.; Government Printing Office).

Kindleberger, Charles P. (1978), *Manias, Panics, and Crashes, a History of Financial Crises* (New York; Basic Books).

Klein, Benjamin (1974), "Competitive Interest Payments on Bank Deposits and the Long-Run Demand for Money," *American Economic Review* (Dec.), LXIV, No. 6, 931–49.

Lucas, Robert (1978), "Asset Prices in an Exchange Economy," *Econometrica* 46, No. 6, 1429–1445.

Macaulay, F. (1938), *The Movements of Interest Rates, Bond Yields, and Stock Prices in the United States Since 1856* (New York; National Bureau of Economic Research).

Miron, J. A. (1985), "Financial Panics, the Seasonality of the Nominal Interest Rate, and the Founding of the Fed," *American Economic Review* 76(1) (March 1986), 125–40.

Mitchell, Wesley C. (1941), *Business Cycles and Their Causes* (Berkeley; University of California Press).

Myers, Margaret (1931), *The New York Money Market*, Volume 1 (New York; Columbia University Press).

Neftci, S. (1979), "Lead-Lag Relations, Exogeneity and Prediction of Economic Time Series," *Econometrica* (January), 104–13.

Noyes, A. D. (1909), *Forty Years of American Finance* (New York; G. P. Putnam's and Sons).

Sargent, T. J. (1971), "Expectations at the Short End of the Yield Curve: An Application of Macaulay's Test," in *Essays on Interest Rates,* edited by J. Guttentag (New York; Columbia University Press and National Bureau of Economic Research).

Shiller, R. J. (1980), "Can the Fed Control Real Interest Rates," in *Rational Expectations and Economic Policy*, edited by Stanley Fischer (Chicago; University of Chicago Press).

Sprague, O. M. W. (1915), "The Crisis of 1914 in the United States," *American Economic Review* (Sept.).

___ (1910), *History of Crises Under the National Banking System* (National Monetary Commission, S. Doc. No. 538, 61st Cong., 2nd sess.).

Tobin, J. (1956), "The Interest-Elasticity of Transactions Demand for Cash," *Review of Economics and Statistics* 38 (August), 241–47.

Waldo, Douglas (1985), "Bank Runs, the Deposit-Currency Ratio and the Interest Rate," *Journal of Monetary Economics* 15(3) (May), 269–278.

White, H. (1981), "Consequences and Detection of Misspecified Nonlinear Regression Models," *Journal of the American Statistical Association* 76, 419–33.

___ (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica* 50, 1–25.

8

Clearinghouses and the Origin of Central Banking in the United States

GARY B. GORTON*

Beginning with Coase's famous essay "The Theory of the Firm," a large literature has developed which seeks to explain why organizations, such as firms, are preferred to a price system for allocating resources.[1] The structure of the U.S. banking industry prior to the existence of the Federal Reserve System is a unique example of such a nonprice allocation system. An essential feature of the banking industry then was the endogenous development of the clearinghouse, a governing association of banks to which individual banks voluntarily abrogated certain rights and powers normally held by firms. Behaving most of the time as the dominant authority in a market-like setting, the clearinghouse was capable of temporarily behaving as a single firm during banking panics. The powers and functions that clearinghouses developed most resembled those of a central bank. In fact, it is almost literally true that the Federal Reserve System, as originally conceived, was simply the nationalization of the private clearinghouse system.

Studying the organization of the pre-1914 banking industry, and, in particular, the role of the clearinghouse, is likely to have implications for assessing the efficiency and uniqueness of contractual arrangements in banking. My note suggests

* The author is Assistant Professor of Finance, The Wharton School, University of Pennsylvania. He gives thanks to the New York City Clearinghouse Association and especially to Gertrude Beck for access to and assistance with their archives.

1. R. H. Coase, "The Nature of the Firm," *Economica* 4 (1937), pp. 386–405.

some working hypotheses about banking industry products and structure and focuses on the New York City Clearinghouse Association (NYCHA) response to banking panics.

8.1. BANK NOTES AND DEMAND DEPOSITS

The clearinghouse emerged with a shift in the relative importance of banking products, products with differing informational and contractual characteristics. The first clearinghouse in the United States, established by New York City banks in 1853, simply created an organized market—a single location where exchange between banks occurred through one other party—the clearinghouse.[2] The rise of demand deposits relative to bank notes, during the latter part of the Free Banking Era (1837 to 1863) necessitated a larger role for the clearinghouse than the market organizer because the demand deposit contract significantly differed from the bank note contract.[3]

Bank notes, small denomination discount bonds, payable in specie on demand at the issuing bank, did not exchange at par outside the bank but at a discount against specie. The system of floating exchange rates between bank notes and specie was possible because secondary markets in bank notes could exist. In fact, the bank note industry consisted of three, sometimes overlapping, types of firms. Banks issued and redeemed notes. Note brokers could earn a return on their investment in information gathering, "making a market" in bank notes, because notes, bought at a discount, could be redeemed at par when "cleared" at the issuing bank. Finally, the prices in these secondary markets were transmitted to agents using the notes in other markets by "monitoring" firms which published bank note reporters and counterfeit detectors.[4] The bank note market revealed information about specific issuing banks so that resources in the bank note industry were allocated by this price system.

A demand deposit, unlike a bank note, is a "double claim" since it is a claim on a specific agent's account at a specific bank. Markets for double claims would be extremely "thin," and it would likely be very costly for brokers to invest in information gathering on every depositor. Also, while in principle checks can circulate by being endorsed, the least costly way to verify the agent-specific dimension of

2. On clearinghouse beginnings see J. S. Gibbons, *The Banks of New York, Their Dealers, The Clearinghouse, and the Panic of 1857* (New York, 1968; reprint of 1859 original); James G. Cannon, *Clearinghouses* (Washington, 1910); Fritz Redlich, *The Molding of American Banking* (New York, 1951), chap. 13.

3. See Redlich, *American Banking*, Part II, p. 3.

4. See William H. Dillistin, *Bank Note Reporters and Counterfeit Detectors, 1826–1866*, Numismatic Notes and Monographs 114 (New York, 1949).

the claim was to "clear" the check quickly. Consequently, private secondary markets in bank checks did not develop. This market was internalized by the banking industry in the form of the clearinghouse, but with the implication that prices did not reveal bank-specific information. In fact, the public exchange rate between checks and specie was fixed at one-to-one. In other words, the demand deposit contract, whereby checks cleared after every transaction, created an information asymmetry between banks and customers because the exchange rates did not fluctuate. Without sufficient price statistics available to depositors to judge the riskiness of banks' deposits, individual banks had an incentive to market deposits with a specie price of less than one, free-riding on the industry. This necessitated a nonprice system to monitor bank performance.[5]

Rather than allocate resources through a price system, the clearinghouse regulated quantities to ensure that the one-to-one exchange rate was accurate. On the one hand, entry to the clearinghouse was screened, and then members were regulated. There were capital requirements, reserve requirements, interest rate restrictions, and ongoing audits and reporting forms to ensure compliance.[6] These efforts were designed to ensure that members did not take advantage of the information asymmetry to reduce the "backing" of their deposits. On the other hand, insofar as deposits were of differing quality, clearinghouses signaled this to the public by requiring members to publish balance sheet items so that the public could adjust their holdings across banks.[7] Threat of expulsion from the clearinghouse was a potent enforcement mechanism.[8]

8.2. THE CLEARINGHOUSE RESPONSE TO PANICS

The U.S. clearinghouse system experienced eight banking panics prior to the creation of the Federal Reserve System. A banking panic occurs with a sudden shift in the perceived riskiness of demand deposits at all banks, leading depositors

5. The argument is developed in greater detail in G. Gorton and D. Mullineaux, "The Joint Production of Confidence: Clearinghouses and the Theory of Hierarchy," 1985, forthcoming.

6. See Cannon, *Clearinghouses.*

7. An important part of the clearinghouses' usual functioning was the investigation of rumors about particular member banks. In response to rumors the clearinghouse, sometimes at the request of the member bank, would audit the bank with its own auditors or auditors hired for that purpose and would then announce the results. There are many examples of this in the New York City Clearinghouse Association, *Clearinghouse Committee Minutes* [hereafter, *Minutes*]. See, for example, April 29, 1873 entry.

8. Member banks were suspended, expelled, and readmitted fairly frequently. For example, the *Minutes* record two member suspensions, six expulsions, four applications for membership declined, four readmissions, and two admissions during the first six years after the clearinghouse was organized.

to demand large-scale transformations of deposits into currency. While the precise variables which can account for panic-causing changes in perceived risk are a matter of debate, information asymmetry creates the possibility of panic. Depositors could not identify bank-specific risk so all banks were vulnerable to runs caused by aggregate events such as increases in business failures.[9] Moreover, in such a setting the failure of individual banks could cause changes in depositors' conditional expectations so that other banks experienced runs. Clearinghouses were institutional responses to both the possibility and the actuality of such information externalities.

When a panic occurred, the structure of the banking industry was radically altered by the metamorphosis of the clearinghouse into a single, firm-like organization uniting the member banks in a hierarchical structure topped by the Clearinghouse Committee. The formation of the new entity was signaled by the first act of the clearinghouse facing a panic, which usually was to suspend the publication of individual bank balance sheet information, publishing instead the aggregate of all members.[10] This was generally accompanied by a joint suspension of convertibility of deposits into currency.[11]

The suppression of bank-specific information, an act completely contrary to the usual functioning of clearinghouses, avoided identifying "weak" banks which might then experience a run which led to runs on other banks. Much more importantly, however, bank-specific information was no longer relevant because banks had joined together in such a way that the aggregate information was, in fact, the appropriate information. The mechanism which united banks was the clearinghouse loan certificate, a liability of the clearinghouse created during panics.

During a panic depositors are demanding that bank portfolios be transformed into securities, the value of which is easily ascertained—namely, specie. Because of the information asymmetry, it is impossible to convince depositors of the value of bank portfolios. The banks themselves, however, were in a position to cope with the problem. The clearing process provided information as did clearinghouse audits and member bank reports. In addition, banks had the

9. See Gary B. Gorton, "Banking Panics and Business Cycles," Philadelphia Federal Reserve Bank, Working Paper, 1984.

10. New York City Clearinghouse Association, *Loan Committee Minutes*, January 30, 1891, June 6, 1893, November 1, 1907; and *Minutes*, November 1, 1907.

11. Suspension of convertibility was avoided during the crises of 1860, 1884, 1895, and 1896. Loan certificates were issued during the crises of 1860 and 1884. In the Panic of 1884 one member did suspend convertibility and was then "suspended from the privileges of the clearinghouse" by unanimous vote (*Minutes*, May 6, 1884). During the crises of 1895 and 1896 the Loan Committee was authorized to issue loan certificates, but no members applied (*Loan Committee Minutes*, December 24–31, 1895, ff., and August 24, 1896).

specialized knowledge to value bank assets. Moreover, banks had an incentive to avoid other members' failures because of the information externalities.

The clearinghouse loan certificate originated during the Panic of 1857 and was used in every subsequent panic through 1914.[12] The process worked as follows. When a panic was imminent or had occurred, the clearinghouse would authorize the issuance of loan certificates. A member bank needing currency to satisfy depositors' demands applied to the clearinghouse's Loan Committee, submitting part of its portfolio as collateral. If acceptable as collateral, certificates were issued amounting to a percentage of the market value of the collateral, that is, bank assets were discounted. The certificates had a fixed maturity of, typically, one to three months, carried an interest charge, and were issued in large denominations.[13] Member banks could use the loan certificates in the clearing process in place of currency, freeing currency for the payment of depositors' claims.

The loan certificates were acceptable in the clearing process not only because they were backed by discounted securities—of greater importance was that loan certificates were claims on the clearinghouse, a joint liability of the members. If a member bank failed and the collateral was worth less than the member's outstanding loan certificates, the loss was shared by the remaining members in proportion to each member's capital relative to the total of all members.[14] The intention of the risk-sharing arrangement, whereby member banks insured each other, was to allow enough currency to be paid out to depositors to signal the soundness of the members while avoiding members' failures.

The coinsurance arrangement, triggered by a panic, did not operate in the usual way markets are thought to operate. The Clearinghouse Committee (and Loan Committee) had a great deal of power in directing the loan certificate process. Not only were the assets submitted as collateral scrutinized by the committee, but the committee had the "power to demand additional security either by an exchange or an increased amount at their discretion."[15] Since the rate of interest on loan certificates and the discount on collateral were the same for all banks (and assets), the power to select and approve collateral and decide on amounts of certificates for individual banks was crucial to the allocation process.

12. See *Minutes*, October 14, 1857 through November 9, 1857.

13. O. M. W. Sprague, *History of Crises Under the National Banking System* (New York, 1968; reprint of 1910 original), pp. 432–33 lists dates of issue, amounts, rate of interest, nature of collateral, and length of issue.

14. The original loan certificate process agreement, *Minutes*, November 21, 1860, does not mention this, though it was made clear during the Panic of 1907 (*Minutes*, October 31, 1907). The Panic of 1907 was apparently the only occasion when members, subsequent to the October 31 resolution could not repay loan certificates.

15. *Minutes*, November 21, 1860.

In addition, the committee apparently had the power to directly allocate the resources of healthy banks to particularly troubled banks. For example, consider this entry in NYCHA minutes, dated October 21, 1907: "The debit balance of the Mercantile Bank having been found to be $1,900,000, it was agreed to extend aid to that bank for the amount of its balance, in addition to the amount already advanced, and the Manager [of the NYCHA] was requested to make requisition on individual banks for the sum of $2,000,000." And there are other examples, as well, of the committee making arrangements for "aid" for members during panics.[16] In general, banks were not allowed to fail during the period of suspension of convertibility, but were expelled from clearinghouse membership for failure to repay loan certificates after the period of suspension had ended.[17]

During banking panics the clearinghouse became a hierarchical structure with the Clearinghouse Committee administering the internal allocation of resources in an attempt to signal to depositors the accuracy of the one-to-one exchange rate for deposit to specie. After a panic, the clearinghouse would revert to its nonpanic form. For the temporary transformation of the clearinghouse to be a viable way for the survival of banking system, the screening and regulatory functions undertaken during nonpanic times had to be successful in limiting the exposure of banks to risk.

8.3. DEPOSIT INSURANCE

During the panics of 1893 and 1907 clearinghouses took the further step of issuing loan certificates, in small denominations, directly to the public.[18] Since this did not involve replacing gold in the clearing process, but instead was the direct monetization of bank portfolios, large amounts of money could be created and issued to the public in exchange for demand deposits. During the Panic of 1893 about $100 million of clearinghouse hand-to-hand money was issued (2.5 percent of the money stock), and, during the Panic of 1907, about $500 million was issued (4.5 percent of the money stock).[19]

Previously, a banking panic was described as an event caused by a shift in the perceived risk of demand deposits at all banks which could happen because

16. See *Minutes*, October 18, 1907, October 21–22, 1907, January 9, 28, 1907, February 1, 1908.

17. *Minutes*, January 30–31, 1908.

18. During the Panic of 1873, the New York City Clearinghouse took an intermediate step by certifying limited amounts of checks as liabilities of the Association. See Sprague, *Crises*, p. 54.

19. John D. Warner, "The Currency Famine of 1893," *Sound Currency*, II (Feb. 15, 1895); A. Piatt Andrew, "Substitutes for Cash in the Panic of 1907," *Quarterly Journal of Economics*, 22 (Aug. 1908), pp. 497–516.

depositors lacked information about bank-specific risk. The loan certificates issued to the public, in exchange for their demand deposits, were acceptable to depositors because they were claims on the association of banks, not just a single bank. Consequently, the exchange of a demand deposit for a loan certificate insured the depositor against individual bank failure. Thus, the problem of bank-specific risk, due to the information asymmetry, was directly addressed.

The loan certificates in the hands of the public were not insurance against the failure of all banks in the association, that is, the failure of the clearinghouse. But, since these claims on the association made bank-specific risk irrelevant to depositors, a secondary market in these claims could and did quickly develop, allowing the risk of clearinghouse failure to be priced. Indeed, a currency premium arose in exchanges of certificates for currency, gradually subsiding until reaching zero, where upon the suspension of convertibility was lifted.[20] This secondary market, reminiscent of bank notes, could exist because of the contractual basis of the loan certificates.

8.4. CONCLUSION

Traditional economic theorizing is strongly biased in favor of markets which operate costlessly through price mechanisms. When applied to banking the paradigm suggests that banking is like any other industry.[21] Yet, by the early twentieth century clearinghouses looked much like central banks. They admitted, expelled, and fined members; they imposed price ceilings, capital requirements, and reserve requirements; they audited members and required the regular submission of balance sheet reports. Finally, they issued money and provided a form of insurance during panics. That such an economic entity should have endogenously arisen in the banking industry suggests important links between the characteristics of the product and institutional and contractual forms of economic organization. While much work remains to be done on these links, the existence of the clearinghouse suggests that private agents can creatively respond to market failure.

20. The currency premia are provided by Sprague, *Crises*, pp. 57, 187, 280–81.

21. For example, see Eugene Fama, "Banking in the Theory of Finance," *Journal of Monetary Economics*, 6 (Jan. 1980), pp. 39–57.

9

The Joint Production of Confidence: Endogenous Regulation and Nineteenth Century Commercial-Bank Clearinghouses

GARY B. GORTON AND DONALD J. MULLINEAUX*

The feasibility of private-market arrangements for the production of money has resurfaced as an important research question (see King 1983 for a review essay). In an early and influential contribution to this literature, Benjamin Klein (1974) emphasized the critical role of consumer confidence in laissez-faire monetary arrangements, and he analyzed "brand names" as potential devices for insuring confidence in private monies.[1] He noted that if monies could not be differentiated, each producer would have incentive to overissue and would do so, unless constrained by some mechanism involving monitoring and control of individual bank behavior. In this regard, Klein notes (p. 441) that "many banks became members of private protective and certifying agencies, which performed

* The authors thank the New York Clearinghouse Association for access to their archives, and Gertrude Beck of the NYCHA for assistance with the archives. They also thank Michael Bordo and members of the staffs of the Federal Banks of Philadelphia and Cleveland for comments on an earlier draft.

1. Vaubel (1977) claims that guarantees, rather than brand name backing, are more likely to be provided in a competitive money-production environment.

some functions similar to present-day central banks." Commercial-bank clearinghouses (CBCHs), for example, utilized regulatory-like tools such as reserve requirements, deposit-rate ceilings, and bank examinations to influence and control the behavior of member institutions.[2]

Based on Klein's analysis, it is somewhat unclear: (1) what motivated commercial banks to voluntarily participate in such arrangements or, (2) why CBCHs were involved in the production of monetary confidence. In this paper, we argue that the evolution of the CBCH reflects an endogenous "regulatory" response to the problems associated with the asymmetric distribution of information in the banking industry. The nature of these information problems was related to the product mix in the banking sector—in particular, to the proportion of demand deposits relative to bank notes. The capacity of "the market" to monitor and control the behavior of bank managers was increasingly eroded as demand deposits came to supplant bank notes during the nineteenth century. The set of actions of the CBCH represent the substitution of hierarchy ("private regulation") for a market-based mechanism of control. That "organizations" may dominate markets as allocation and control devices is hardly a new idea (Coase 1937, Williamson 1975, and Stiglitz 1985).

In section 9.1, we discuss the importance of the banking product mix during the nineteenth century from the viewpoint of information costs. Section 9.2 describes the role of the CBCH as a monitor/supervisor which provides valuable "screening" services to both member banks and the public. Section 9.3 examines the behavior of the CBCH during financial panics. In response to the unusual information costs associated with a panic, the CBCH increased the amount of private regulation. The CBCH then reverted to its simpler organizational form following the conclusion of a panic. Private regulation declined and the role of "the market" as a control mechanism increased. Section 9.4 concludes.

9.1. BANK NOTES, BANK DEPOSITS, AND INFORMATION COSTS

Bank notes involved a contract between the bearer and the bank to redeem the face value of the note in specie at the bank. The specie value of a bank note to a seller accepting it in exchange was simply the expected value of a bank's specie promise less the costs of collecting specie at that bank. Even if the expected specie value of a note was par, the collection costs drove a wedge between the

2. Gorton (1985b) and Timberlake (1984) have called still more explicit attention to the strong similarities between the activities of nineteenth century CBCHs and today's Federal Reserve System. Neither of these authors explored in depth the reasons why clearinghouses took on regulatory-like activities, however.

par value of a note and its value in exchange for goods. This wedge created an incentive for note-broker businesses to form offering to exchange bank notes for gold or the notes of other banks at discounted rates. Brokers could profit by collecting specie at par at the issuing bank. Such firms indeed did form, and a secondary market in bank notes emerged. The size of the discounts quoted on notes presumably varied with the geographic distance to the issuing bank, the perceived riskiness of that institution and the quantity of counterfeit notes of that institution believed to be in circulation relative to the total issue (Gorton 1986). In "bank-note reporters," brokers published information on counterfeits along with current quotes on various notes.

Secondary market makers also had strong incentives to monitor the quality of the assets backing bank notes since they collected specie in bulk as the source of their profitability. Their price quotations in turn revealed their information to buyers and sellers of bank notes. Indeed, merchants commonly consulted bank-note reporters in reaching judgments about the exchange values of particular bank notes. Competition among note brokers and publishers of note reporters presumably enhanced the information quality of these price signals (Dillistin 1949, White 1895). To the extent that brokers returned notes to the bank of issue, they also performed a clearing and collection function. Thus, while bank notes typically exchanged for goods and services at a discount, the overall variability in these discounts was constrained by the self-correcting responses of banks, note brokers, and consumers to the recurring signals provided by the secondary market in bank notes.

A demand deposit, unlike a bank note, is both a claim on a bank and on an agent's account at that bank. This complicates the information required to price a check claim on that deposit. In an exchange mediated by check, the seller of goods must consider (1) whether the check writer has sufficient funds for the check to be collected; (2) whether the check writer's bank can exchange for specie; and perhaps (3) whether his own bank can exchange for specie at par. While the identity of a buyer "doesn't matter" with use of a bank note (in the absence of counterfeits), a check-based transaction is agent-specific with respect to risk.

The contractual characteristics of demand deposits accordingly increased the transactions costs associated with this product. These costs in turn precluded the development of a secondary market in claims on such deposits. Such a market would require pricing agent-specific claims on a bank. It would prove extremely costly for specialist note brokers to acquire information on the reliability of individuals as well as banks. Yet such information is necessary to price such a claim since the agent issuing a check can overdraw his balances.

Banks were better able than note brokers to handle the information-related disadvantages of checks. Banks could delay specie payment on checks, for instance, until after checks were collected. This required an accounting system,

but such a system was a necessary adjunct to producing demand deposits. Also, banks could assume that some proportion of the checks collected would be held as deposits rather than paid in specie. These deposits could fund income-producing assets. Brokers could not offer deposit-type accounts, at least not without the risk of being considered a bank, and therefore having to submit to chartering requirements and perhaps other regulations.

The contractual differences between bank deposits and notes effectively precluded brokers from competing with banks in the collection of deposits. Accordingly, no "secondary market" in check claims was formed. As a result, the information production of the note brokers concerning the "quality" of individual banks became increasingly less available as the volume of deposits increased relative to notes. Holders of bank liabilities therefore could monitor bank behavior only in a direct and costly fashion.

Banks in the cities had a larger proportion of their liabilities as deposits than as bank notes as early as the late eighteenth century. The Bank of New York reported in 1791 that it had 50 percent more deposits than notes outstanding. Data became regularly available in the 1830s and show a fairly steady decline in the notes/deposits ratio. In New York state, for example, the notes/deposit in ratio was 1.2 in 1837, 0.74 in 1847, and 0.31 in 1857 (Redlich 1951). Nationally, the trend was less pronounced. The ratio fell from 0.85 in 1835, to 0.79 in 1845, and to 0.67 in 1860 (see *Historical Statistics of the United States*, p. 995).

Given their informational disadvantages, it may seem curious that deposits came to dominate bank notes rather early in the century, even before the establishment of the first CBCH. But demand deposits do possess certain well-known advantages over bank notes. They are less subject to theft, for example. In addition, writing checks avoids the cost of making change and provides proof of payment. Another less commonly recognized feature of using checks rather than notes to make payments is that checks exchanged against currency or goods and services in local markets at a fixed price. While the specie price of a particular bank's notes could vary dramatically over time and space, deposits, when acceptable to sellers in transactions, exchanged at par in local transactions. But if deposits were to prove viable in exchange, some mechanism for providing confidence in performance by banks was necessary. This was especially the case since a uniform exchange rate for deposits created incentives for banks to "cheat" by backing deposits with inferior assets. There was no secondary market to "reveal" such behavior as there was with bank notes.

The formation of the CBCH not only reduced the costs of clearing checks, it solved the information problem created by the missing market, by *internalizing* the secondary market in a unique organizational form. With the CBCH, the apparent defects of the demand deposit product could be turned into distinct advantages.

9.2. THE CLEARINGHOUSE AS A MONITOR/MANAGER

The CBCH was not initially formed to deal with resource allocation problems which markets handle poorly. Its function was to economize on the costs of check clearing. Prior to the New York CBCH formation in 1853, commercial banks collected checks and other instruments by a daily exchange and settlement with each other bank. Once the clearinghouse formed, the exchange was made with only one party—the clearinghouse itself. Gibbons (1859) estimates that for New York City banks the cost of "conducting this vast amount of business did not exceed eight thousand dollars a year," which constituted roughly 0.02 percent of deposits in the New York CBCH at the end of 1854.

While the clearinghouse was organized to produce a simple product, check-clearing, it was also capable of producing a by-product—information. When demand deposits dominate bank notes, banks have an exploitable information advantage over their customers concerning the quality of bank liabilities. Banks face incentives to back deposits with high-yielding, risky assets. Customers want to obtain information about the true quality of bank deposits, but face free rider problems. The direct statement of the bank lacks credibility since a "bad" bank has no incentive to reveal its true condition. Customers would clearly gain if some form of credible supervisor monitored the quality of bank liabilities and disseminated relevant information. Such a supervisor would need enforcement powers to correct contract deviations. The supervisor, in other words, would act as a substitute for the price system; hierarchy (authority) would replace the market.

Such a system would be implemented if it were in the welfare interests of the banks as well as their customers. The gain to an individual bank from industry supervision is identical to that for employees in a firm: colleagues can shirk only at a higher cost. Even though workers see compulsion as costly, they are better off in a number of circumstances by accepting it (Stiglitz 1975). This becomes more true as shirking by colleagues reduces the return to an individual worker or increases his risk. When deposits dominate, banking is characterized by just such a condition, since shirking by one bank can lower the return to another directly. A "bad" bank's failure or suspension, for example, would induce bank customers to monitor the quality of their own bank's liabilities.[3] The cheapest way to monitor was to exercise the deposit contract. But if large numbers of customers chose to monitor at once (a bank run), even a "good" bank ran a substantial risk of failure. This externality problem strengthened the demand for supervision, other things equal. The "best" banks would favor monitoring even aside from externalities since disclosure of their status may allow them to capture "ability rents."

3. Suspension was a temporary default on the contract to exchange bank liabilities for specie.

There are strong reasons in favor of quality measurement by the banks themselves. Bank measurement need occur only once per measurement period, for example, but customer measurement involves a great deal of duplication. In addition, bankers possess comparative advantages in judging the quality of the assets backing deposits.

The CBCH was well positioned to provide monitoring and supervision services to the banking industry. The form of the New York clearinghouse, embodied in its 1854 constitution, included a number of aspects similar to institutions commonly identified today as providing screening services, mainly educational institutions. The clearinghouse required, for example, that member institutions satisfy an admissions test (based on certification of adequate capital), pay an admissions fee, and submit to periodic exams (audits) by the clearinghouse. Members who failed to satisfy CBCH regulations were subject to disciplinary actions (fines) and, for extreme violations, could be expelled.

Expulsion from the clearinghouse was a clear negative signal concerning the quality of bank's liabilities. It suggested that in the collective judgment of the banking community, the probability of nonperformance in the exchange process by the expelled bank was uncomfortably high. The ability of the CBCH to audit a member's books (to measure quality) at any moment provided strong incentives for prudent behavior by each bank and thus strengthened the credibility of the CBCH signals.[4] Moreover, without access to the clearinghouse a bank had to clear its checks in the more costly manner used prior to the existence of the CBCH. Consequently, expulsion was a potent enforcement threat.

The CBCH also increased the value of other information signals. Each bank in New York City was required by law to publish on each Tuesday morning a statement showing the average amount of loans and discounts, specie, deposits, and circulation for the preceding week. Banks were also required to publish quarterly statements of condition. The existence of the CBCH prevented banks from publishing inaccurate statements and from engaging in excessive "window dressing" of balance sheets.[5]

The advantage of the CBCH organization was such that within a decade a large number of new local clearinghouses were formed. These typically

4. Gibbons writes: "With knowledge of these facts (debits in excess of specie balances for a sustained period), the Committee visits the bank, and investigates its affairs. If they are found to be hopelessly involved, it is suspended from the exchange at the Clearing House—a last blow to its credit" (pp. 319–20). Dismissal from the clearinghouse required only a majority vote.

5. "It was only when the Clearing House *records* were brought to such perfection as to give the means of analysis and test beyond dispute, that the positive integrity of those statements could be guaranteed to the public" (Gibbons 1859, p. 325). The CBCH would also investigate rumors about the states of particular member banks. In response to rumors, the CBCH would audit the bank and publish the results. There are many examples of this in the New York City Clearinghouse Association, Clearinghouse Committee *Minutes* (hereafter, *Minutes*).

organized along lines similar to the New York CBCH, but some extended their roles beyond that of monitoring to regulating bank behavior. The Buffalo and Sioux City clearinghouses set interest-rate ceilings on deposits which could be paid by member banks (Cannon 1910).

The New York CBCH did not employ fixed reserve requirements as a supervisor-enforced constraint on members until 1858, when a 20 percent "coin requirement" was established against "net deposits of every kind" (Hammond 1957, p. 713). Reserve requirements were also soon thereafter established in Philadelphia. The reserve requirement did *not* apply against circulating *notes*. The CBCH also monitored the extent to which members purchased or borrowed specie from external sources to meet claims. Member banks were, in effect, under implicit contract to the CBCH to avoid "excessive liability management."[6]

These activities of CBCHs served to enforce the fixed local exchange rate of one-to-one between specie and demand deposits. By credibly supervising member bank activities and by reducing the costs of clearing checks, CBCHs helped demand deposits become the preferred bank product on the liability side. But one problem remained: how would bank liability holders monitor the monitor?

9.3. THE CLEARINGHOUSE DURING BANKING PANICS

The behavior of CBCHs was consistent with a hierarchical form of organization focused principally on supervisory kinds of activities. But, while the costs of member-bank "cheating" were raised by the CBCH, it could not eliminate all incentives to cheat. Indeed, by raising the public's perception of the quality of the "average" bank, the CBCH raised the benefit of cheating along with the cost. There remained some incentive, therefore, for bank customers to engage in their own monitoring of bank behavior. A banking panic may be seen as an instance of customer monitoring. Exercising the deposit contract's option feature en masse represents a cheap way for bank customers to monitor the ability of their bank to perform, and, in effect, to monitor the monitoring of the CBCH.

Banking panics were large-scale attempts by bank customers to convert deposits into specie or currency. While the precise causes of banking panics remain a point of dispute, it seems clear that, because of the information asymmetry created by demand deposits, depositors had to rely on aggregate or nonbank-specific information to assess the riskiness of deposits. Increases in

6. "A positive principle, or rule of financial government, has been demonstrated by this action of the Clearing House on the city banks—that is, the restriction of loans, by the necessity of maintaining a certain average of coin *from resources within the bank*. Borrowing from day to day will no longer do. It cannot be concealed." (Italics original, Gibbons 1859, p. 321.)

business failures or the failure of a single large financial firm could cause depositors to "run" on all banks seeking, in a single act, to withdraw deposits and measure the performance of their individual banks and, implicitly, the performance of the CBCH (Gorton 1984).

From a bank's point of view, there are potentially large costs to such measurement by its customers. The customers can only be convinced of the value of demand deposits if the banks can transform them into specie or currency. With bank notes, the secondary market signaled the value of bank portfolios in an efficient manner. But without a secondary note market, bank claim holders had to rely on nonmarket methods of evaluation. In part because of the high cost of obtaining information on the quality of bank loans, this portion of a bank's assets can be deemed illiquid. If the sale of such illiquid assets is required to meet depositors' demands, then a bank may incur substantial losses. In other words, the excessive measurement by customers which occurs during a panic effectively makes illiquidity the same as insolvency.

With costless, full information, the banking system would never face problems during panics because bank assets could easily be transformed into any other desired securities. But in that case there would never be a panic to start with because depositors would never need to monitor. With an information asymmetry, banks would value some mechanism which allowed for their assets to be transformed into some other security in such a way as to signal to depositors their value. The CBCH provided such a mechanism by inventing a new security, the clearinghouse loan certificate.

The first issue of clearinghouse loan certificates occurred during the panic of 1857; they were issued in every subsequent panic through 1914. The process was straightforward: a policy committee of the CBCH First authorized the issuance of loan certificates. Member banks needing specie or currency to satisfy customers' demands could then apply to the clearinghouse loan committee for certificates. Borrowing banks were charged interest rates varying from 6 to 7 percent and were required to present "acceptable collateral" to be "discounted" by the CBCH. The loan certificates had a fixed maturity of, typically, one to three months. The important feature of the certificates was that member banks could use the loan certificates in the clearing process in place of currency, freeing currency for the payment of depositors' claims.[7]

The mechanism of the loan certificate produced a more hierarchical organizational form of the CBCH during panics than existed otherwise. Indeed, during panics when the loan certificate process was operating, the CBCH behaved much like an integrated firm allocating resources by hierarchical decision. In fact,

7. The dates of issue, amounts issued, rate of interest, and nature of collateral can be found in the *Report* of the U.S. Treasury, 1914, p. 589. In the pre–Civil War, "bills receivable, stocks, bonds, and other securities" were acceptable. Also see Sprague (1910), pp. 432–33.

the loan certificates were claims on the clearinghouse, a joint liability of the member banks. If a member bank with outstanding loan certificates failed, the loss (in excess of the value of pledged collateral) was shared by the remaining members of the CBCH.[8]

The loan certificate process in effect internalized the missing market within a hierarchical form. While depositors faced an information asymmetry, the banks themselves were in a position to cope with this problem. The clearing process itself provided information on members, as did clearinghouse audits and member bank reports. Also, banks had the specialized knowledge to value bank assets. Most importantly, individual banks had an incentive to lower the probability of other members' failures because of the information externalities. This meant in practice that no member banks were allowed to fail during a period of panic. Instead, members were expelled from clearinghouse membership for failure to repay loan certificates after the panic had clearly ended and their failure would result in weaker externality effects.

The loan certificate process was available to all members, and consequently, is accurately described as a coinsurance arrangement. But this meant that resources had to be allocated to members, even those which the CBCH perhaps knew would certainly fail, in the interests of all members. Since the interest rate on loan certificates and the discount on collateral did not vary over banks or assets, the central decisions of selecting and approving collateral, and deciding on amounts of certificates were *quantity* decisions made by the CBCH. Moreover, the CBCH could, at its discretion, demand additional security and requisition aid for particularly troubled banks.[9] The CBCH clearly possessed a great deal of control. It regulated bank behavior substantially during a panic.

8. In New York the first explicit record of how loan certificates were to function, *Minutes,* November 21, 1860, does not mention this. It was made clear during the Panic of 1907 (*Minutes*, October 31, 1907) which was apparently the only occasion when, after the panic, members (two banks) could not repay loan certificates. However, during the first panic the CBCHs faced after formation, a particularly lucid statement of this was adopted by the Boston CH (October 15, 1857). The agreement is quoted in Redlich (1951), p. 159.

9. In Boston the original 1857 agreement included the following:

> And it is further agreed . . . that the Clearing House Committee may at any moment call upon any bank for satisfactory collateral security, for any balance thus paid in bills instead of Specie; and each Bank hereby agrees with the Clearing House Committee, and with all and each of the other Banks to furnish immediately such security when demanded.
>
> Quoted in REDLICH (1951), p. 159.

In New York the CH Committee had the "power to demand additional security either by an exchange or an increased amount at their discretion" (*Minutes*, November 21, 1860). But beyond this was power to directly allocate resources by making requisitions on individual banks (*Minutes*, October 21, 1907). Also, see *Minutes*, October 18, 1907; October 21–22, 1907; January 9, 28, 1907; February 1, 1908.

Another managerial decision in which the CBCH became involved was when and whether to suspend the right of deposit convertibility, that is, to suspend the option feature of deposit contracts. Suspension amounted to default on the deposit contract, and was a violation of banking law. Nevertheless, suspension occurred on eight occasions during the nineteenth century.[10] In banking panics after 1853, the CBCH played the central role in deciding whether and when suspension was appropriate.[11] Suspension signals that the CBCH believes further liquidation of bank assets to acquire currency or specie is not in the welfare interests of either the suspending banks or their customers (Gorton 1985a).

The transformation of the CBCH into a single firm-like organization during panics was signaled by suspending the weekly publication of individual bank statements, and instead, publishing the weekly statement of the clearinghouse itself.[12] In this way, the clearinghouse avoided identifying weak banks. But, more importantly, with the loan certificate process at work, the aggregate information was the appropriate information. Also, the CBCH did not publish the identity of banks borrowing through the loan-certificate process. Cannon (1910, p. 90) reports that "attempts on the part of the business community were made in vain to discover what banks had taken out in certificates."

For this organizational structure to be successful, the amount of currency released from use in the clearing process through use of loan certificates had to be large enough to signal to depositors that the one-to-one deposit exchange rate was, in fact, correct. But the amount of currency released was limited, and so, during the panics of 1893 and 1907, the clearinghouses directly monetized bank portfolios by issuing loan certificates, in small denominations (as low as 25 cents), directly to the public. This allowed all the banks' assets to be monetized, if needed.[13]

Depositors were willing to accept loan certificates in exchange for demand deposits (rather than currency) because the loan certificates, being claims on the

10. Suspension of convertibility occurred during August 1814, Fall 1819, May 1837, October 1857, September 1873, July 1893, and October 1907. Suspension also occurred in the 1860s though this was not related to a major banking panic as in the other cases. Loan certificates were issued during every panic after the formation of the CBCH, including 1860 and 1884. During the crises of 1895 and 1896 the New York City CBCH authorized the issuance of loan certificates, but no member banks applied (*Loan Committee Minutes*, December 24–31, 1895; August 24, 1896).

11. For example, the Marine National Bank was punished for acting on its own by unilaterally suspending in May, 1884 (*Minutes*, May 6, 1884). The New York CBCH avoided suspension during the Panic of 1884.

12. E.g., *Loan Committee Minutes*, January 30, 1891; June 6, 1893; November 1, 1907; and *Minutes*, November 1, 1907.

13. Gorton (1985b) computes that the U.S. money stock temporarily increased in this way by $2\frac{1}{2}$ percent during the Panic of 1893 and by $4\frac{1}{2}$ percent during the Panic of 1907.

CBCH, insured depositors against individual bank failure. In this way, the problem of bank-specific risk arising from the information asymmetry was solved, leaving only the risk that the CBCH would fail. But the circulating loan certificates were neither bank- nor agent-specific, so a secondary market could and did quickly develop, allowing the risk of CBCH failure to be priced. This secondary market served as an index of confidence. Initially, a currency premium existed in exchanges of certificates for currency.[14] Over the period of suspension, it gradually subsided until reaching zero, whereupon suspension was lifted. In this way, a market signal was sent from depositors to CBCHs.

During banking panics, the CBCH was operating a miniature capital market, allocating resources by nonmarket means for the benefit of the collective of firms. But once the period of suspension was over, the CBCH reverted to its more limited organizational form. Only by reverting back to the more limited organizational form could the CBCH restore the proper incentives for banks to jointly monitor each other on a continuous basis.

Suppose that once the more hierarchical form of organization had been adopted during a panic, the CBCH did not revert back to its more limited form. Then individual banks, knowing that the loan certificates were available, would have an incentive to make riskier loans since each would believe that the risk could be spread over the other members through the loan certificate process. Clearly, this would not be viable. During the period of suspension when the risk pooling arrangement was in effect, however, banks have incentives to make more risky loans, free-riding on the CBCH. No mention of such a problem appears in the archives of the New York Clearinghouse Association or other sources. The problem apparently didn't exist because member banks had no funds to make new loans. During panics banks attempted to liquidate loans of existing customers to generate cash. If a member did engage in making riskier loans, however, it was exposed to the risk that the maturity of the loans would be longer than the suspension period, making free-riding less likely. Also, the CBCH required daily reporting of all balance-sheet changes during a panic period.

Only by reverting back to the more limited organizational form did individual banks have the incentives to monitor each other. The externalities from individual bank cheating provided the incentives, and the resulting monitoring made it possible for the panic-form of the CBCH to be effective since the risk exposure of the members had been limited during nonpanic times. Consequently, the changing organizational form and degree of regulation of the CBCH was an integral part of the production of demand deposit services. In the absence of a market to monitor product quality, bank firms were required to jointly produce "confidence" in deposits, but this required a delicate balance between hierarchy and maintenance of market incentives.

14. See Sprague (1910), pp. 57, 187, 280–81.

9.4. CONCLUSION

Analysis of the CBCH system focuses attention on the issue most critical to the discussion of competitive banking: the ability of "the market" to control the behavior of bank managers. Hayek (1976) and White (1984) have argued that market forces are capable of controlling banks, and consequently preserving confidence in the system, provided that bank liabilities are convertible into some outside money. Klein (1974) has emphasized the role of brand names in establishing and maintaining confidence concerning convertibility. We have argued, however, that, because of information asymmetries, the market's capacity to control bank behavior depends on the banking product mix. In particular, the rising ratio of deposits to bank notes during the nineteenth century resulted in (1) increased monitoring costs for bank customers, and (2) more significant externality problems among banks. The CBCH, originally formed as a simple collective to reduce the costs of collecting checks, became involved in monitoring activities and established mechanisms of managerial control. In fact, the CBCH "regulated" bank behavior.

Our analysis provides a more complete and consistent explanation for the role of private institutions such as the CBCH in the creation of monetary confidence, which has been noted by Klein (1974), Timberlake (1984), and Gorton (1985).[15] It also suggests that the conclusions of Hayek (1976) and White (1984) concerning the efficacy of markets as control mechanisms in banking may be valid only under certain conditions concerning information costs and monitoring technologies.

REFERENCES

Cannon, James G. *Clearing Houses* (U.S. National Monetary Commission). Washington: Government Printing Office, 1910.

Coase, Ronald H. "The Nature of the Firm." *Economica* 4 (November 1937), 386–405.

Dillistin, William. *Bank Note Reporters and Counterfeit Detectors, 1826–1866.* New York: American Numismatic Society, 1949.

Gibbons, James S. *The Banks of New York, Their Dealers, the Clearinghouses, and the Panic of 1857.* New York, 1859.

Gorton, Gary B. "Banking Panics and Business Cycles." Mimeographed. The Wharton School, University of Pennsylvania, 1984.

____. "Bank Suspension of Convertibility." *Journal of Monetary Economics* 15 (March 1985a), 177–94.

____. "Clearinghouses and the Origins of Central Banking in the U.S." *Journal of Economic History* 42 (June 1985b), 277–84.

15. Mullineaux (1987) analyzes role of a different private institution, the Suffolk Bank System, in maintaining confidence in bank notes in New England during the mid-nineteenth century.

____. "Inside Money and Contracting Technologies: An Empirical Study." Mimeographed. The Wharton School, University of Pennsylvania, 1986.

Hammond, Bray. *Banks and Politics in America.* Princeton: Princeton University Press, 1957.

Hayek, Friedrich A. Von *The Denationalisation of Money.* London: Institute of Monetary Affairs, 1976.

King, Robert. "On the Economics of Private Money." *Journal of Monetary Economics* 12 (July 1983), 127–58.

Klein, Benjamin. "The Competitive Supply of Money." *Journal of Money, Credit, and Banking* 6 (November 1974), 423–53.

Mullineaux, Donald J. "Competitive Monies and the Suffolk Bank System." *Southern Economic Journal* 53 (April 1987), 884–98.

Redlich, Fritz. *The Molding of American Banking.* New York: Hafner Publishing Company, 1951.

Sprague, Oliver M. W. *History of Crises under the National Banking System* (U.S. National Monetary Commission). Washington: Government Printing Office, 1910.

Stiglitz, Joseph E. "Incentives, Risk and Information: Notes toward a Theory of Hierarchy." *Bell Journal of Economics* 6 (Autumn 1975), 552–79.

____. "Credit Markets and the Control of Capital." *Journal of Money, Credit, and Banking* 17 (May 1985), 133–52.

Timberlake, Richard H., Jr. "The Central Banking Role of Clearinghouse Associations." *Journal of Money, Credit, and Banking* 16 (February 1984), 1–15.

Vaubel, Roland. "Free Currency Competition." *Weltwirtschafliches Archiv* 113 (September 1977), 435–59.

White, Lawrence H. *Free Banking in Britain: Theory Experience and Debate, 1800–45.* Cambridge: Cambridge University Press, 1984.

Williamson, Oliver. *Markets and Hierarchies: Analysis and Antitrust Implications.* New York: Free Press, 1975.

10

Bank Panics and the Endogeneity of Central Banking

GARY B. GORTON AND LIXIN HUANG*

10.1. INTRODUCTION

The most important function of a central bank is to provide liquidity to the banking system in times of crisis. The classic work on central banking, Lombard Street, by Walter Bagehot, published in 1877, offered the advice that in times of panic the central bank (Bank of England) should lend freely and continue to pay out currency (Bagehot, 1877). At the time Lombard Street was published, there was no central bank in the US and yet the private arrangement of banks in the US clearinghouse system had already discovered Bagehot's precepts and was acting on them. In this chapter we argue that the lender-of-last-resort function of "central banking" arose endogenously through the formation of state contingent bank coalitions, such as clearinghouses, which provided liquidity during banking panics.

In the model we propose, central banking emerges endogenously as a response to the banking system's problems of asymmetric information and concomitant moral hazard. In some banking systems these problems can lead to banking panics. But, these banking panics are not irrational manifestations of multiple equilibria. Rather, these episodes represent depositors monitoring their banks, which are vulnerable to moral hazard problems in certain states of the world. With the information asymmetry, such panics may involve inefficiencies because banks may be mistakenly liquidated. Banks cannot honor the demands

* Thanks to Franklin Allen, Eslyn Jean-Baptiste, Michael Bordo, John Boyd, Charles Calomiris, Ed Green, Richard Kihlstrom, Holger Mueller, Ben Polak and to seminar participants at New York University, the Yale Banking Conference, and the Federal Reserve Bank of Cleveland Conference on the Origins of Central Banking for comments and suggestions. Huang acknowledges the financial support of the Hong Kong RGC Competitive Earmarked Research Grant (City U 1252/03H).

of all depositors; there is not enough liquidity in the banking system. Efficiency can be improved in two ways. First, banks can be more accurately identified, so that only those banks in bad states are liquidated. Secondly, liquidity can be created which, as we show below, mitigates the problem of moral hazard. The industrial organization of the banking system is crucial in determining whether these improvements are operable. We show how central banking arose endogenously as a by-product of the interaction between the industrial organization of banking and the problems emanating from asymmetric information.

Specifically, we study three different organizational forms of the banking industry: a system with small independent unit banks; a system with a few highly branched and well-diversified big banks; and a system with a bank coalition. The unit banking system is the least efficient, because it suffers from severe asymmetric information problems, due in part to the fact that these banks are not diversified. Costly economy-wide liquidations following banking panics are the only way to forestall moral hazard. The big bank system is more efficient for two reasons. First, diversification alleviates the asymmetric information problem so that mistaken bank runs can be avoided. Second, big banks can self-monitor by closing branches to improve the quality of assets. The self-monitoring mechanism enables big banks to send credible signals to depositors that incentives to engage in moral hazard have been removed. Once depositors' confidence is restored, bank runs are stopped. The bank coalition system partially replicates the big bank system in certain states of the world through state contingent coalition operations, including mutual monitoring and liability pooling. However, ownership and property rights of individual banks give rise to incentive compatibility constraints that prevent coalitions from fully replicating big banks.

The implications of the model are consistent with banking history. A comparison of the US and Canadian banking experiences from the middle of the 19th century is a particularly instructive example of the importance of industrial organization in banking and its relation to central banking. Haubrich (1990), Bordo et al. (1994, 1995), and White (1984), among others, study the drastic contrast between these two systems. During the period 1870 to 1913, Canada had a branch banking system with about 40 chartered banks, each extensively branched, while at the same time the US had thousands of banks that could not branch across state lines. The US experienced panics, while Canada did not.[1] There were high failure rates in the US and low failure rates in Canada. Thirteen Canadian banks failed from 1868 to 1889, while during the same period hundreds of banks failed in the US (see the Comptroller of the Currency, 1920). During the Great Depression, there were few bank failures in Canada, but the Canadian banking system did shrink by about the same amount as in the US (see

1. Calomiris and Gorton (1991) identify six panics in the US prior to 1865, seven during the National Banking Era.

White, 1984). Overall, the Canadian banking system survived the Great Depression with few effects, while in the US, which had enacted the Federal Reserve Act in 1914, the banking system collapsed. Canada's central bank came into being in 1935, well after the Great Depression.

Associated with the likelihood of bank panics is the prevalence of private arrangements among banks. In the US, for example, where panics were not infrequent, the private clearinghouse system developed over the course of the 19th century (see Cannon, 1910; Sprague, 1910; Timberlake, 1984; Gorton, 1984, 1985; Gorton and Mullineaux, 1987; and Moen and Tallman, 2000, among others). During a banking panic member banks were allowed to apply to a clearinghouse committee, submitting assets as collateral in exchange for "clearing house loan certificates," which is a form of private money issued by bank coalitions. The loan certificates were the joint liability of the clearinghouse, not the individual bank. During the Panics of 1873, 1893, and 1907 the clearinghouse loan certificates were issued directly to the banks' depositors, in exchange for demand deposits, in denominations corresponding to currency.[2] If the depositors would accept the certificates as money, then the banks' illiquid loan portfolios would be directly monetized. In this way, a depositor who was fearful that his particular bank might fail was able to insure against this event by trading his claim on the individual bank for a claim on the portfolio of banks in the clearinghouse. This lender-of-last-resort function was the origin of deposit insurance.

Bank coalitions are also not unique to the US. There are many examples of bank coalitions forming on occasion in other countries as well (see Cannon, 1910 for information on the clearinghouses of England, Canada, and Japan). We mention a few examples. According to Bordo and Redish (1987), the Bank of Montreal (founded in 1817) emerged very early as the government's bank performing many central bank functions. The pattern of the Bank of Montreal (and earlier precursors like the Suffolk Bank in the US) in which the bank coalition is centered on one large bank, is quite common. Similarly, in Germany the Bankhaus Herstatt was closed June 26, 1974. There was no statutory deposit insurance scheme in Germany, but the West German Federal Association of banks used $7.8 million in insurance to cover the losses. Germany is a developed capitalist country where deposit insurance is completely private, being provided by coalitions of private banks that developed following the Herstatt crisis of 1974 (see Beck, 2001).

2. The amount of private money issued during times of panic was substantial. During the Panic of 1893 about $100 million of clearinghouse hand-to-hand money was issued (2.5 percent of the money stock). During the Panic of 1907, about $500 million was issued (4.5 percent of the money stock). See Gorton (1985).

The paper proceeds as follows. In Section 10.2 we present a simple model of a banking system that is then analyzed in subsequent sections. Our first step is to analyze two polar cases using the model. The first case is a banking system with small independent unit banks (Section 10.3) and the second is a system of large, well-diversified, branched banks (Section 10.4). Neither of these systems literally represents reality, though they come close to the experiences of some countries. The US historically has been a system of small independent unit banks and when private clearinghouses were in existence, not all banks were members.[3] The system of large branched banks, the other polar case, does resemble many of the world's banking systems, such as Canada. In Section 10.5 we consider the system with small independent unit banks that can form a coalition in the event of a banking panic. Section 10.6 concludes. Proofs of the propositions can be found in the appendix of the paper on SSRN or NBER Working Paper #9102.[4]

10.2. THE MODEL

There are three dates, 0, 1, and 2 in the model economy and two types of agents: consumers/depositors and bankers. Bankers are unique in having the ability to locate risky investment opportunities. Also, only banks can store endowments (i.e., provide the service of safekeeping).

There is a continuum of bankers. Each banker has capital β and a measure one of potential depositors. Each bank has access to a riskless storage technology and to a risky investment technology. The fraction of the portfolio invested in the riskless storage technology is α; this investment will be referred to as reserves. The remaining fraction $1 - \alpha + \beta$ is invested in the risky technology. Investments in the risky projects have to be made at date 0, and the returns are realized at date 2. The return to a unit (of endowment good) invested in the risky project is $\tilde{\pi} + \tilde{r}$, that is, there is a systematic component, $\tilde{\pi}$, and an idiosyncratic component, $\tilde{r}$, to the return. So, the state of the macroeconomy is indicated by $\tilde{\pi}$, while the bank's individual prospects are indicated by $\tilde{r}$. We assume that $\tilde{\pi}$ is uniformly distributed in the interval $[\pi_{\mathrm{L}}, \pi_{\mathrm{H}}]$ and $\tilde{r}$ is uniformly distributed in the interval $[0, 2\mathrm{M}]$. For future reference, the probability density function of $\tilde{\pi}$ will be referred to as A, where $A \equiv 1/(\pi_{\mathrm{H}} - \pi_{\mathrm{L}})$.

At date 1, information about the date 2 return is realized, but there is asymmetric information between bankers and depositors. Depositors observe the realized state of the macroeconomy (π), but they do not observe the realized

3. Some banks were too far away to be members. Rural banks and banks in smaller cities did not have formal clearinghouse arrangements.

4. See http://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=17513 or http://papers.nber.org/papers/w9102.

state of their bank's idiosyncratic return (r). Each banker knows his own bank's state (r), and observes the realizations of other banks' idiosyncratic shocks at date 1. Idiosyncratic shock realizations at date 1 are not verifiable among banks, but realized cash flows at date 2 are verifiable. So, to be clear, banks cannot write contracts with other banks contingent on idiosyncratic shocks at date 1. At date 0, we assume that banks' capital β and choice of reserve level α are observable and verifiable.

There is a moral hazard problem in that bankers have an opportunity to engage in fraud at date 1. Fraud is socially wasteful. If a banker engages in fraud, he gets a proportion f of the return, i.e., $f(\pi + r)$, where f is strictly less than 1. The remaining amount, $(1-f)(\pi + r)$, is wasted and depositors receive nothing. Projects can be liquidated at date 1, yielding a constant return of Q, regardless of the state of the project.

Depositors have a subsistence level of 1. Their utility function is:

$$u(c_1, c_2) = \begin{cases} c_0 + c_1(1+\varepsilon_1) + c_2(1+\varepsilon_2) & \text{if} \quad c_0 + c_1 + c_2 \geq 1, \\ -\infty & \text{if} \quad c_0 + c_1 + c_2 < 1, \end{cases}$$

where c_0, c_1, and c_2 are consumptions at dates 0, 1, and 2, respectively. ε_1 and ε_2 represent depositors' preference for later consumption. We assume $\varepsilon_2 > \varepsilon_1 > 0$ and they are both very small such that they can be ignored in the following analysis. The depositors' utility function implies that they will always wait until date 2 to withdraw if they believe their deposits are safe. However, they will withdraw at date 1 if they anticipate that there is any chance that their bankers are going to engage in fraud. Depositors deposit in a single bank.

Because of their utility functions, depositors need to be assured that their claim will be worth 1 unit and banks can satisfy this need. Implicitly, individual banks can diversify to this extent. Gorton and Pennacchi (1990) show that uninformed consumers/traders with uncertain consumption demands prefer to transfer wealth intertemporally with riskless claims. A better arrangement for these consumers could be claims on a diversified bank that are always worth 1 unit (i.e., so that there is no private information that informed traders could take advantage of). We do not explicitly incorporate all this here. Rather, in the model here the structure of preferences dictates the type of claim that banks will offer depositors: the bank must offer the right to withdraw deposits at face value at date 1, i.e., a demand deposit contract.

We assume that bankers are risk neutral and they get the entire surplus from investment. In addition, we assume the following:

Assumption 1. $(1+\beta)(1-f)(\pi_{\mathrm{L}}+M) < 1$. This assumption assures that there is a potential moral hazard problem. Suppose a banker invests all of his assets in the risky project, and the economy turns out to be in the worst possible state (π_{L}) at date 1. Consider the banker with the mean return $\pi_{\mathrm{L}} + M$. If he engages

in fraud, he will receive $f(1+\beta)(\pi_L+M)$. If he does not engage in fraud, his payoff will be $(1+\beta)(\pi_L+M)-1$. The assumption $(1+\beta)(1-f)(\pi_L+M)<1$ implies that the banker has an incentive to engage in fraud.

Assumption 2. $\pi_L > Q > f(\pi_H+2M)$. In words, there is a dead weight loss if liquidation or fraud occurs. If fraud does not occur, then the value of a risky project is greater than the liquidation value, Q, even if the project is in the lowest possible state. If fraud occurs, then the value of a risky project is less than the liquidation value even if the project is in the highest possible state.

Assumption 3. $(\pi_L+\pi_H)/2+M>1>Q$. This assumption says that, ex ante, a risky project is more efficient than riskless storage, if there is no liquidation or fraud. However, if liquidation or fraud happens, then a risky project is dominated by investment in riskless storage.

Assumption 4. $(1+\beta)Q>1$. That is, if depositors withdraw from their bank at date 1, then their deposit contract can always be honored.

Assumption 5. A risky project is indivisible when liquidation occurs. Although at date 0, a banker can choose how much to invest in a risky project, at date 1 all the assets in a risky project must be liquidated if liquidation occurs.

The essential ingredients of the model are the moral hazard problem and the information asymmetry. Fraud, the assumed moral hazard in this model, has historically been the most common reason for bank failure. The Comptroller of the Currency (1873), reporting on the Panic of 1873, wrote that all the bank failures during the panic were due to "the criminal mismanagement of their officers or to the neglect or violation of the national-bank act on the part of their directors" (p. xxxv). A century later, the Comptroller of the Currency (1988b) reported that:

> The study found insider abuse in many of the failed and rehabilitated banks during their decline. Insider abuse—e.g., self-dealing, undue dependence on the bank for income or services by a board member or shareholder, inappropriate transactions with affiliates, or unauthorized transactions by management—was a significant factor leading to failure in 35 percent of the failed banks. About a quarter of the banks with significant insider abuse also had significant problems involving material fraud. (p. 9)

For purposes of the model, it is important that there be a moral hazard problem, but it is not essential that the problem be fraud. Any one of a number of moral hazard problems would suffice. Fraud, however, is a realistic and significant problem.

Since a banker may have an incentive to engage in moral hazard in certain states of the world, actions needs to be taken to stop them. Specifically in this model, we make the following definition.

Definition. *Monitoring* means to prevent a bank from engaging in fraud.

There are different ways to prevent fraud (or monitor the banks). The simplest way is to take the assets away from the bankers.

Definition. A *bank run* is an event in which a large number of depositors, fearing the banker engaging in fraud, withdraw their funds at date 1. A bank panic is an event in which many banks suffer from bank runs.

Because of the problem of information asymmetry, there can be "good" runs and "bad" runs. Good runs prevent the moral hazard problems; bad runs force banks that are not going to suffer from moral hazard problems to liquidate their projects. According to Assumption 2, good runs are efficient while bad runs are inefficient. If the information asymmetry problem can be alleviated, then bad runs might be avoided. This generates the demand for a lender-of-last-resort.

Definition. A *lender-of-last-resort* is an institution which provides liquidity to banks so that they do not have to liquidate their projects.

Note that liquidity provision has broader meanings than cash injection. For example, if an institution can provide insurance for a bank, then a run can be stopped. A more interesting example is that the lender-of-last-resort can save a bank by delivering a convincing signal that the bank is in good state. In other words, alleviating information asymmetry is also a way to provide liquidity. This is the point we want to emphasize in this paper.

Bankers can commit to not engaging in moral hazard by holding reserves. The higher the level of reserves, the lower the probability of a bank run. However, ex post, if the state of the economy is good at date 1, then it would have been better to have invested reserves in risky projects. The bankers' task at date 0 is to choose an optimal reserve level, α (the fraction of bank assets held in the riskless storage technology). This is the only choice variable. The optimal reserve choice depends on whether bank branching is allowed and on the interaction between the bankers. We interpret branching restrictions and different interactions between the bankers as different banking systems. We consider three basic forms of organization, two polar cases and one intermediate case. The first case is a system of many small independent unit banks. The next is a system of large, well-diversified banks, and the last is a system of small unit banks that can form a coalition in certain states of the world. Below, we proceed to solve the bankers' optimization problem under the different organizations of the banking industry, examining the reserve level, banking stability, and social welfare under each system.

10.3. THE SYSTEM OF INDEPENDENT UNIT BANKS

The first banking system we examine is one in which there are many small, independent unit banks. That is, implicitly the banks are small so they are undiversified. This is because they have no branches and they do not interact with

each other ex ante or ex post (they are independent). This system characterizes those periods of US history, for example, where banks were not allowed to branch and where they did not form explicit or implicit coalitions. We will call this banking system the "unit bank" system.

Unit banks are "small" in the following sense: a banker in charge of a unit bank can only manage one risky project. Implicitly, we imagine that banks are spatially separated so that risky projects have the idiosyncratic risk of the individual bank's location. A banker only has the expertise in managing the project in his local region. The assumption also implies that at date 1, the project of a banker cannot be transferred to another banker, who lacks the skill to manage it. In other words, a project involves a relationship specific investment that cannot be transferred.

We solve the bankers' optimization problem by backward induction. First, given a unit bank's choice of reserve level, α, we characterize the states in which bankers will have incentives to engage in moral hazard and, hence, depositors will withdraw their deposits. Second, we will calculate the bankers' optimal choice of reserve level, α, at date 0.

At date 1, depositors receive the signal about the state of the macroeconomy, π; they do not observe the realization of their bank's idiosyncratic shock, r. Because their utility functions are kinked and they will get minus infinity if consumption is less than one, they do not care about the likelihood of their bank engaging in fraud, but only consider whether there is any chance of this occurring. They, therefore, assume that $r = 0$ and check whether their banker has an incentive to engage in fraud. Suppose a banker has reserves α and the realized state of the macroeconomy is π. If the banker does not engage in fraud, his payoff will be $\pi(1 + \beta - \alpha) + \alpha - 1$. If he engages in fraud his payoff will be $\pi f(1+\beta-\alpha)$, since he cannot steal anything from the reserves. The banker has incentive to engage in fraud if and only if $\pi f(1+\beta-\alpha) > \pi(1+\beta-\alpha) + \alpha - 1$, or $\pi < (1-\alpha)/\left(\left(1-f\right)(1+\beta-\alpha)\right)$. If such an incentive exists $\left(\text{i.e. } \pi < (1-\alpha)/\left(\left(1-f\right)(1+\beta-\alpha)\right)\right)$, depositors withdraw all their savings. Since all the depositors receive the same macroeconomic information and all the banks are, from their viewpoint, homogeneous, if one bank suffers from a run, there are runs on all the other banks. Therefore, a panic occurs.

At date 0, anticipating what will happen in different states of the world at date 1, bankers choose the optimal reserve level to maximize their expected payoff. On the one hand, bankers want to maximize investment in the risky projects because this is more profitable, but on the other hand, they want to avoid being prematurely liquidated in a banking panic at date 1. If bankers hold reserves such that $\alpha \geq \alpha_{\max}^{\mathrm{U}} \equiv (1-(1+\beta)(1-f)\pi_{\mathrm{L}})/(1-(1-f)\pi_{\mathrm{L}})$, then they have no incentive to engage in the moral hazard even if the economy is in the lowest state.

Therefore, bankers solve the following optimization problem at date 0:

$$\begin{aligned}
&\text{Max}_{\alpha} \int_{\pi_{\text{L}}}^{\pi^r} \left[\alpha + (1+\beta-\alpha)Q - 1\right] \mathrm{d}F(\tilde{\pi}) \\
&\quad + \int_{\pi^r}^{\pi_{\text{H}}} \left[\alpha + (1+\beta-\alpha)(\pi + M) - 1\right] \mathrm{d}F(\tilde{\pi}) \\
&\text{s.t. } \pi^r = \frac{1-\alpha}{(1-f)(1+\beta-\alpha)}, \alpha \in \left[0,\, \alpha_{\text{max}}^{\text{U}}\right].
\end{aligned}$$

PROPOSITION 1. *There exists a unique optimal reserve level* $\alpha \in \left[0,\, \alpha_{\text{max}}^{\text{U}}\right]$ *that solves the bankers' optimization problem.*

(The proof of Proposition 1, and all other proofs can be found in the longer version of the paper on SSRN or the NBER Working paper.)

The purpose of a panic is to monitor the bankers, to prevent them from engaging in fraud. The panic is not irrational; it is not motivated by externalities due to actions of other depositors when there is a sequential service constraint. Rather, the panic is related to the macroeconomy, which may create incentives for bankers to engage in moral hazard. The fear of not being able to satisfy subsistence should the banker engage in moral hazard, a kind of extreme risk aversion, causes the depositors' withdrawals. However, not all bankers will engage in moral hazard. The problem is that depositors do not know which bankers have high idiosyncratic shock realizations and which have low idiosyncratic shock realizations. Depositors liquidate all banks because of information asymmetry. Bankers hold high reserves to avoid being liquidated, but ex ante this is inefficient.

10.4. THE BIG BANK SYSTEM

At the other extreme from a banking system composed of many independent unit banks is a system where banks are large and heavily branched, or well diversified. We call this the "big bank" system. Most banking systems in the world are closer to this system than to the system of independent unit banks, discussed above.

Because a big bank is well diversified, it has a portfolio of assets that has a realized return of $\pi + M$ at date 1. In other words, a big bank's return is the systematic return plus the diversified idiosyncratic mean return, M. This is the essential point, namely, that the idiosyncratic risk is diversified away, implicitly by virtue of the bank's size via branching. Consequently, at date 1, the state of macroeconomy is sufficient information for assessing the state of a big bank. As a result, depositors know for sure whether a big bank is going to engage in moral hazard or not and they never run a big bank wrongly.

In addition to more transparent information, a second benefit is that a big bank has the flexibility to liquidate individual projects. By liquidating projects with low idiosyncratic returns, a big bank implements "self-monitoring" and improves the quality of assets. More importantly, since branch closure is public information, depositors know that a big bank's situation has been improved. If a big bank can convince depositors that its incentive to engage in moral hazard has been eliminated by self-monitoring, depositors' confidence can be restored and they will allow the big bank to continue the non-liquidated projects till completion.

Since the risky projects have the same liquidation value Q in spite of their idiosyncratic returns, a big bank will always liquidate those projects that have the lowest realized idiosyncratic returns. Suppose the big bank is to liquidate a fraction x of the risky projects. It will liquidate those projects with realized idiosyncratic returns, r, in the interval $[0, x2M]$. The average return on the remaining, i.e., nonliquidated, $(1-x)$ fraction of projects is $\pi + (x2M + 2M)/2 = \pi + (1+x)M$. In order to make a commitment not to engage in fraud, the big bank has to liquidate a fraction x of the risky projects such that:

$$\begin{aligned}&\alpha + (1+\beta-\alpha)xQ + (1+\beta-\alpha)(1-x)(\pi + (1+x)M) - 1\\ &\geq (1+\beta-\alpha)(1-x)(\pi + (1+x)M).\end{aligned}$$

This is a quadratic inequality, which admits a solution in the interval $[0, 1]$,[5]

$$x = \frac{Q-(1-f)\pi - \sqrt{(Q-(1-f)\pi)^2 - (4M(1-f)(1-\alpha)/1+\beta-\alpha) + 4M(1-f)^2(\pi+M)}}{2M(1-f)}.$$

Anticipating what will happen in different states of the world at date 1, the banker who owns a big bank chooses the optimal reserve level to maximize the expected payoff at date 0. If the reserve level is higher than $\alpha^{B}_{max} \equiv (1-(1+\beta)(1-f)(\pi_L + M))/(1-(1-f)(\pi_L + M))$, then he has no incentive to engage in moral hazard even if the economy is in the lowest state. The date 0 optimization problem can be written as

$$\begin{aligned}&\text{Max}_{\alpha} \int_{\pi_L}^{\pi^r} \left\{\alpha + (1+\beta-\alpha)xQ + \left[(1+\beta-\alpha)(1-x)(\pi + (1+x)M) - 1\right]\right\}\\ &\quad dF(\tilde{\pi}) + \int_{\pi^r}^{\pi_H} \left[\alpha + (1+\beta-\alpha)(\pi + M) - 1\right] dF(\tilde{\pi})\end{aligned}$$

5. Since depositors can only observe how many branches, but not which branches are closed, they form consistent with the Big Bank's action on the equilibrium path. It is easy to characterize depositors' off-equilibrium path beliefs. For example, depositors always believe that a Big Bank closes the branches with lowest idiosyncratic returns and use this belief to check the Big Bank's incentive.

Table 10-1. BANK BALANCE SHEET ITEMS FOR CANADA AND THE US, 1870–1919

	1870–79	1880–89	1890–99	1900–09	1910–19
Canada					
Loan/assets	0.717	0.706	0.696	0.722	0.640
Securities/assets	0.013	0.021	0.071	0.087	0.110
Debt/equity	1.458	1.914	2.796	4.232	6.876
United States					
Loan/assets	0.487	0.563	0.589	0.546	0.567
Securities/assets	0.253	0.169	0.117	0.164	0.168
Debt/equity	1.826	2.334	2.620	4.184	5.352

SOURCE: Table 4 of Bordo et al. (1995) (based on US Comptroller of the Currency, Annual Reports, and Curtis, 1931).

$$\text{s.t. } \pi^r + M = \frac{1-\alpha}{(1-f)(1+\beta-\alpha)},$$

$$\alpha \in \left[0,\ \alpha^B_{\max}\right],$$

$$x = \frac{Q-(1-f)\pi - \sqrt{(Q-(1-f)\pi)^2 - (4M(1-f)(1-\alpha)/1+\beta-\alpha) + 4M(1-f)^2(\pi+M)}}{2M(1-f)}$$

PROPOSITION 2. *The above objective function is strictly concave in* α. *There is a unique optimal reserve level,* $\alpha \in \left[0,\ \alpha^B_{\max}\right]$ *that solves the big bank's optimization problem.*

To emphasize, note that in the big bank system banks may experience withdrawals at date 1, but they do not fail because of their ability of "self-monitoring." This is the major difference from the unit banking system, in which bank panics cause bank failures. In the big bank system, although some of the projects might be liquidated and branches closed, the system can survive. The unit banking system cannot survive panics.

In broad outlines, the distinction between the big bank system and the system of small independent unit banks corresponds to the difference between the Canadian and US systems. As mentioned above, the Canadian system generally displayed fewer failures and no panics. In addition, as table 10.1 makes clear, Canadian banks held fewer reserves (in the form of securities) and, correspondingly, they made more loans per asset dollar.

After 1920, the comparison is also stark. By 1920 the private clearinghouse system in the US that functioned as a lender-of-last-resort was gone, having been replaced by the Federal Reserve System. In Canada, the bank merger movement, from 1900 to 1925, reduced the number of banks and resulted in a small number of banks with large branch networks. Prior to the merger movement, Canadian banks were branched, but there were many more banks. The post-merger movement banking system in Canada is clearly the big bank system. The comparison

between the two systems during this period is the subject of Bordo et al. (1994), who emphasize the fact that between 1920 and 1980 there was one bank failure in Canada, in contrast to hundreds and thousands in the US, particularly during the Great Depression. There were no banking panics in Canada, though the reduction in deposits during the Great Depression was of similar magnitude, as noted above.

10.5. BANK COALITIONS

The above two polar cases, the unit banking system and the big bank system, can be thought of as representative benchmarks. In this section, we introduce the possibility of a bank coalition, i.e., a state contingent agreement between banks. The discussion of bank coalitions will follow the US clearinghouse experience, briefly described above, but the argument is more general, as discussed below.

The basic idea for the coalition is as follows. The failure of individual small unit banks as a result of bank runs at date 1, despite holding high levels of reserves, can be improved upon if the small banks can replicate, at least partially, the performance of a big bank. Diversification allows big banks to alleviate the information asymmetry problem. In addition, big banks can "monitor" themselves by liquidating part of their portfolio, in the face of withdrawals, to boost the depositors' confidence. For small banks to attempt to replicate the performance of a big bank, a mechanism that achieves these two functions must be invented.

Credibility of a coalition is established by a signal of its solvency; the signal is the coalition's act of issuing claims backed by all member banks to depositors in exchange for individual bank deposits. These claims, the loan certificates, are supported by a sharing rule that combines assets and liabilities at date 1 and provides incentives for the member banks with high idiosyncratic shock realizations to monitor member banks with low idiosyncratic shock realizations. "Monitoring" means to prevent member banks from engaging in moral hazard, by liquidating these banks or subsidizing them. The internal workings of the coalition are not observable to depositors, so they will not accept the loan certificates unless they believe that the coalition's behavior is incentive compatible. In equilibrium depositors' beliefs will be consistent with the behavior of the coalition. We now turn to providing the details.

10.6. THE SETTING WITH BANK COALITIONS

Suppose that there are small independent unit banks at date 0. They are prohibited from forming a big bank. (For example, banks are prohibited from branching across state lines.) Without forming a big bank, however, unit banks can get together to form a coalition by reaching an agreement about their individual

capital and reserve levels at date 0. The coalition stipulates date 1 state contingent rules indicating which banks to be liquidated and how to share liabilities among the remaining non-liquidated banks. Because the idiosyncratic shocks are not verifiable, and thus not contractible, the coalition has no power to force its members to comply with the rules and the member banks are free to quit at any time they want. In other words, coalition rules have to be incentive compatible. Depositors cannot observe whether the rules have been carried out or not at date 1. They can only observe whether the coalition liquidates some of the member banks and combines the assets and liabilities of the remaining member banks.

The sequence of events at date 1 begins with depositors observing the realized state of the macroeconomy and deciding whether to withdraw their deposits or not. Then the banks decide whether to trigger the operation of the coalition. We define the coalition and the operation of the coalition as follows:

Definition. The bank coalition is an agreement between member banks at date 0 about the following issues to maximize the total payoffs to its member banks:[6]

(i) Bank reserve levels, α, at date 0.
(ii) A date 1 state-contingent rule, $P(\alpha, \pi)$, indicating when the coalition is to operate ($P = 1$) or not operate ($P = 0$). If $P = 0$, then banks act as unit banks. (The contingency, in fact, will be a panic; this is shown below.)
(iii) If the coalition is set into operation, then the coalition applies two rules: a liquidation rule $L(\alpha, \pi, r)$, which is a mapping from $[0, 2M]$ to $\{1, 0\}$, indicating whether a member bank with idiosyncratic shock r is to be liquidated ($L = 1$) or not liquidated ($L = 0$); and a debt transfer rule, $D(\alpha, \pi, r)$,[7] which is a mapping from $[0, 2M]$ to R^+, indicating the liability reallocated to a non-liquidated member bank with idiosyncratic shock r. Deposits in non-liquidated banks are replaced with loan certificates, which are debt claims of the coalition, backed by all the assets of all the member banks.

The operation of the coalition is intended to achieve two goals. First, by liquidating some of the member banks the coalition tries to inform depositors that

6. For simplicity, we do not go into the details how decisions are made inside the coalition. We assume that the internal organization of the coalition is equivalent to assuming the existence of a coalition decision maker who is independent of any of the member banks and maximizes the total payoffs to all member banks.

7. Although r is not verifiable, D is verifiable at date 2. Moreover, the coalition needs to prevent member banks revealing their r by showing depositors their $D(\alpha, \pi, r)$. We can imagine that the coalition takes out a note "You owe the coalition $D(\alpha, \pi, r)$" and asks the banker for his signature. In this way, only the coalition holds the verifiable contracts, which specify all non-liquidated banks' liabilities $D(\alpha, \pi, r)$.

"mutual monitoring" has started and the non-liquidated banks are in relatively more sound states. This partially alleviates the asymmetric information between the banks and depositors. Second, by pooling the liabilities the coalition quells depositors' concern over banks' idiosyncratic states and convinces depositors that the coalition as a whole is healthy and good banks will provide liquidity to bad banks.

10.7. EQUILIBRIUM WITH BANK COALITIONS

Suppose at date 1 the systematic macroeconomic state, π, is realized. Define $r^* \equiv (1-\alpha)/\left(\left(1-f\right)(1+\beta-\alpha)\right) - \pi$. We have $f(1+\beta-\alpha)(\pi+r) > \alpha + (1+\beta-\alpha)(\pi+r) - 1$ for $r < r^*$ i.e., only banks with $r < r^*$ have incentives to engage in fraud. If these banks act as unit banks (i.e., the coalition does not operate, $P(\alpha, \pi) = 0$), the total payoff to all the banks is

$$\int_0^{r^*} f(1+\beta-\alpha)(\pi+r)\mathrm{d}F(\tilde{r}) + \int_{r^*}^{2M} \left[\alpha + (1+\beta-\alpha)(\pi+r) - 1\right]\mathrm{d}F(\tilde{r}).$$

If the coalition operates ($P(\alpha, \pi) = 1$), the liquidation rule and the sharing rule are carried out. Under any given coalition rules, the coalition as a whole can be either solvent or insolvent at date 2. If it is solvent, the total of payoff cannot exceed $\int_0^{2M} [\alpha + (1+\beta-\alpha)(\pi+r) - 1]\,\mathrm{d}F(\tilde{r})$, which can be reached when none of the member banks engages in fraud. If it is insolvent, the total payoff cannot exceed $\int_0^{2M} f(1+\beta-\alpha)(\pi+r)\,\mathrm{d}F(\tilde{r})$, which can be reached when all member banks engage in fraud. Therefore, the maximum total payoff a coalition can achieve is

$$\begin{aligned}
&\max\left\{\int_0^{2M} [\alpha + (1+\beta-\alpha)(\pi+r) - 1]\,\mathrm{d}F(\tilde{r}),\right.\\
&\left.\int_0^{2M} f(1+\beta-\alpha)(\pi+r)\,\mathrm{d}F(\tilde{r})\right\}. \text{ Since}\\
&\int_0^{r^*} f(1+\beta-\alpha)(\pi+r)\,\mathrm{d}F(\tilde{r})\\
&+\int_{r^*}^{2M} [\alpha + (1+\beta-\alpha)(\pi+r) - 1]\,\mathrm{d}F(\tilde{r})\\
&\geq \max\left\{\int_0^{2M} [\alpha + (1+\beta-\alpha)(\pi+r) - 1]\,\mathrm{d}F(\tilde{r}),\right.\\
&\left.\int_0^{2M} f(1+\beta-\alpha)(\pi+r)\,\mathrm{d}F(\tilde{r})\right\},
\end{aligned}$$

it is better not to carry out the coalition rules voluntarily.

Depositors are rational and they understand that the coalition is not going to operate without a bank panic. So they run the banks to trigger the state contingent operation of the coalition.

Once the depositors run the banks, the coalition has to operate to convince the depositors that it will exert its monitoring and coinsurance functions, the following proposition presents the details.

PROPOSITION 3. *Suppose that at date* 1 $\pi < (1-\alpha) / \left(\left(1-f\right)(1+\beta-\alpha)\right)$ *and, consequently, depositors run the banks. Define*

$$x^*(\alpha, \pi) \equiv \max\left\{0, \min\left\{1, \frac{1-\alpha-\pi\left(1-f\right)(1+\beta-\alpha)}{M\left(1-f\right)(1+\beta-\alpha)} - 1\right\}\right\}.$$

The coalition operates, i.e., $P(\alpha, \pi) = 1$. It applies the liquidation rule, setting $L(\alpha, \pi, r) = 1$ (i.e., liquidation) for banks with idiosyncratic shocks $r \in [0, x*(\alpha, \pi) 2M]$ and pays these bankers $\alpha + (1+\beta-\alpha) Q - 1$. For non-liquidated banks, the coalition reallocates liabilities according to the members type, r: $D(\alpha, \pi, r) = \alpha + \left(1-f\right)(1+\beta-\alpha)(\pi+r)$; loan certificates backed by all non-liquidated banks are issued to replace deposits in these banks. All member banks comply with coalition rules.

The proposition shows how the coalition behaves as a lender-of-last-resort by monitoring and by providing insurance.[8] Monitoring corresponds to liquidating bad banks, those with the worst idiosyncratic shock realizations ($r \in [0, x*(\alpha, \pi) 2M]$). These banks would have engaged in fraud. The insurance comes from the transfers implemented among the non-liquidated banks ($r \in [x*(\alpha, \pi) 2M, 2M]$). These banks are assigned new debt obligations according to $D(\alpha, \pi, r) = \alpha + \left(1-f\right)(1+\beta-\alpha)(\pi+r)$. Their original debt, i.e., face value of the demand deposits, was one. Note that banks with $r < (1-\alpha) / \left(\left(1-f\right)(1+\beta-\alpha)\right) - \pi$ have their liabilities reduced, i.e., $D(\alpha, \pi, r) < 1$, so they no longer have incentives to engage in fraud. This is efficient because the continuation values of these projects are worth more than the liquidation value if they are immune to fraud. Member banks with $r > (1-\alpha) / \left(\left(1-f\right)(1+\beta-\alpha)\right) - \pi$ have their liabilities increased, i.e., $D(\alpha, \pi, r) > 1$, so they are taxed to pay the subsidy to the low r banks. Banks with high idiosyncratic shock realizations cannot be taxed too much, or they will engage in fraud. The transfers of the debt obligations must satisfy the budget

8. Here again, for expositional purposes, we omit the characterization of the off-equilibrium path beliefs held by depositors when they observe that the number of banks liquidated by the coalition is different from what the liquidation rules stipulate. The most reasonable belief is that depositors always believe the coalition liquidates banks with lowest idiosyncratic returns. It is easy to check that such a belief does not allow the coalition to deviate from the liquidation rule and the sharing rule.

constraint $\int_{x^*}^{2M} D(\alpha, \pi, r)\, \mathrm{d}F(\tilde{r}) = 1$. This limits how much liquidity the coalition can provide and, therefore, determines the point at which member banks are liquidated.

The banking panic creates an externality for banks that would not engage in the moral hazard problem, the "good" banks. Without a panic, they would have no incentive to monitor the banks that are going to engage in fraud, the "bad" banks. Because depositors do not know the idiosyncratic states of each bank and bad banks can always mimic good banks, good banks cannot renege on their responsibilities by quitting the coalition. Facing the prospect of being liquidated, they are forced to monitor the bad banks by liquidating the worst ones and providing liquidity to the others via liability sharing.

There is a critical difference between how the coalition and the big bank deal with panics. The difference has to do with the difference between the ownership and property rights in these two systems. The banker of a big bank (implicitly) hires branch managers to manage branches for him, and he gets the entire surplus. We do not need to consider the branch managers' incentives because the branch manager has no property rights over his branch.[9] A coalition member cannot be forced to operate his bank in a certain way, nor can he be involuntarily separated from his assets. Consequently, when a big bank closes a branch, it gets $\alpha + (1 + \beta - \alpha) Q - 1$ after paying off the branch depositors and uses this amount as additional reserves. These additional reserves change the incentives of the big bank. But, the coalition cannot increase reserves in this way because member banks have the property rights and hence control of their assets; they are free to quit the coalition. In addition, while the big bank "monitors" itself, the coalition works through "mutual monitoring." Non-liquidated good banks need to bribe/subsidize non-liquidated bad banks to keep them from engaging in fraud. This restricts the coalition's liquidation rule and the sharing rule because each banker has to be promised a payoff at least equal to what he can get from quitting the coalition or staying and engaging in fraud. Otherwise the coalition would have more freedom to set these rules and act more like a big bank.

At date 0, each bank must decide whether to join the coalition and the coalition must determine the optimal reserve level α. The optimal reserve for the coalition is the solution of the following problem:

9. The banker of a big bank is the owner and has the cash flow rights. Even though a manager can engage in fraud, it does not mean he can reap the benefit of doing so, because the realized cash flows go to the banker first before they are redistributed to the managers. In addition, since the banker has the full control, it is easy for him acquire evidence and bring a manager to the court in case the manager engages in fraud without his agreement

$$\begin{aligned}
\text{Max}_{\alpha} & \int_{\pi_L}^{\pi^r} \left[\alpha + (1+\beta-\alpha)xQ + (1-x) \times (\pi + (1+x)M)\right] dF(\tilde{\pi}) \\
& \quad + \int_{\pi^r}^{\pi_H} \left[\alpha + (1+\beta-\alpha)(\pi + M)\right] dF(\tilde{\pi}) - 1 \\
\text{s.t.} \quad & \pi^r = \frac{1-\alpha}{(1-f)(1+\beta-\alpha)}, \\
& x = \max\left\{0, \min\left\{1, \frac{1-\alpha-\pi\left(1-f\right)(1+\beta-\alpha)}{M\left(1-f\right)(1+\beta-\alpha)} - 1\right\}\right\}, \\
& \alpha \in \left[0, \alpha_{\max}^{U}\right].
\end{aligned}$$

PROPOSITION 4. *The coalition's objective function is strictly concave in* α. *There is a unique optimal reserve level,* $\alpha \in \left[0, \alpha_{\max}^{U}\right]$, *that solves the coalition's optimization problem. At date* 0, *every bank strictly prefers to join the coalition.*

The coalition system is an intermediate case between the unit banking system and the big bank system. When the macroeconomy is in the good state, the coalition system is the same as the Unit Bank system. Contingent on banking panics following a negative systematic shock, the coalition system is triggered and mutual monitoring and insurance take place. The coalition partially replicates the big bank. The unique feature associated with the coalition is that when a panic occurs, it suspends convertibility and issues certificates. This feature is important because it is a commitment made to depositors that the nonliquidated member banks will not engage in fraud and it provides incentives for member banks to monitor and insure each other. The role of suspension of convertibility here is quite different from a coordination device used to eliminate Pareto-dominant equilibria in other models (e.g., Diamond and Dybvig, 1983).

10.8. COMPARING THE DIFFERENT BANK SYSTEMS

We have studied three different banking systems: the independent unit banking system, the big bank system, and the bank coalition. In this section, we compare these systems in terms of welfare. Keep in mind that, on the one hand, holding reserves is inefficient because the risky project earns a higher return. But, on the other hand, holding fewer reserves means a higher chance of a panic and project liquidation.

PROPOSITION 5. *The unit banking system holds more reserves than the coalition system, which, in turn, holds more reserves than the big bank system. The big bank system is more efficient than the coalition system, which is more efficient than the independent unit banking system.*

In the unit banking system, if depositors monitor banks by withdrawing, then the bank panic results in all banks being liquidated. Independent unit banks cannot monitor each other, nor do they have (private) deposit insurance like the coalition system. Banks in the unit banking system can only resort to excess reserves to avoid the ex post losses from forced liquidations. The big bank has two advantages. First, diversification eliminates the information asymmetry problem. And second, it can close branches and use the proceeds as reserves to alter its incentives to engage in fraud—self-monitoring and liquidity provision. Such advantages allow the big bank to invest more in the risky projects and hold less reserves. The coalition system lies between the unit banking system and the big bank system. State contingent monitoring and co-insurance provide banks in the coalition with a way to survive panics if they are solvent. However, because property rights in the coalition do not allow it to completely replicate the big bank, mutual monitoring and insurance is not as efficient as self-monitoring and liquidity improving, and banks in the coalition have to hold more reserves than banks in the big bank system.

10.9. DISCUSSION

We studied the relation between the industrial organization of banking and banking panics. Banking panics occur in systems of small unit banks. Panics result from depositors monitoring/liquidating banks in a setting where some banks are more likely to be engaging in moral hazard, but the depositors do not know which banks are the more likely because of asymmetric information. Banking systems with large, well-diversified, banks are more efficient because diversification alleviates asymmetric information problem. In addition, branch closure as a publicly observable self-monitoring mechanism allows big banks to improve the quality of assets and restore depositors' confidence. When branching is not allowed, the lender-of-last-resort functions, including money creation, monitoring, and deposit insurance arose from private arrangements among banks. Small banks form bank coalitions to monitor members and provide insurance to depositors. Banking panics play a crucial role in making such private bank coalitions work. They impose an externality on member banks so that they are forced to commit to pool resources and liquidate some members.

Why did government central banks replace private bank coalitions? In the above analysis, there is no obvious rationale for the government to step in and provide the lender-of-last-resort function unless the government has much more power than private agents, more resources than private agents, or there are costs to panics that have not been considered. Gorton and Huang (2003) consider the above model, but include a transactions role for bank liabilities. A panic disrupts the role of bank liabilities as a medium of exchange. They argue that in this context the government may be able to improve welfare with deposit insurance.

REFERENCES

Bagehot, W., 1877. Lombard Street. Charles Scribner's Sons, New York.

Beck, T., 2001. Deposit Insurance as a Private Club: Is Germany a Model? World Bank mimeo.

Bordo, M., Redish, A., 1987. Why did the Bank of Canada emerge in 1935? Journal of Economic History 47, 405–17.

Bordo, M., Redish, A., Rockoff, H., 1994. The US banking system from a northern exposure: stability versus efficiency. Journal of Economic History 54, 325–41.

Bordo, M., Rockoff, H., Redish, A., 1995. A comparison of the stability and efficiency of the Canadian and American banking systems, 1870–1925. NBER Historical Working Paper #67.

Calomiris, C., Gorton, G., 1991. The origins of banking panics: models, facts, and bank regulation. In: Hubbard, G. (Ed.), Financial Markets and Financial Crises. University of Chicago Press, Chicago, pp. 109–73.

Cannon, J.G., 1910. Clearing Houses, US National Monetary Commission, 61st Congress, 2nd Session, Doc. No. 491. Government Printing Office, Washington DC.

Comptroller of the Currency, 1873. Annual Report. Government Printing Office, Washington, DC.

Comptroller of the Currency, 1920. Annual Report. Government Printing Office, Washington DC.

Comptroller of the Currency, 1988b. Bank Failure: An Evaluation of the Factors Contributing to the Failure of National. Government Printing Office, Washington DC.

Diamond, D., Dybvig, P., 1983. Bank Runs, deposit insurance, and liquidity. Journal of Political Economy 91, 401–19.

Gorton, G., 1984. Private clearinghouses and the origins of central banking. Federal Reserve Bank of Philadelphia Business Review (January–February), pp. 3–12.

Gorton, G., 1985. Clearinghouses and the origins of central banking in the US. Journal of Economic History 45, 277–83.

Gorton, G., Huang, L., 2003. Banking Panics and the Origin of Central Banking. Chapter 5. In: Altig, D., Smith, B. (Eds.), Evolution and Procedures of Central Banking. Cambridge University Press, Cambridge, pp. 181–219.

Gorton, G., Mullineaux, D., 1987. The Joint production of confidence: Endogenous regulation and the 19th century commercial bank clearinghouse. Journal of Money, Credit, and Banking 19, 457–68.

Gorton, G., Pennacchi, G., 1990. Financial intermediaries and liquidity creation. Journal of Finance 45, 49–72.

Haubrich, J., 1990. Nonmonetary effects of financial crises: lessons from the great depression in Canada. Journal of Monetary Economics 25, 223–52.

Moen, J., Tallman, E., 2000. Clearinghouse membership and deposit contraction during the panic of 1907. Journal of Economic History 60, 145–63.

Sprague, O.M.W., 1910. History of Crises Under the National Banking System. Government Printing Office, Washington, DC.

Timberlake, R., 1984. The central banking role of clearinghouse associations. Journal of Money, Credit, and Banking 16, 1–15.

White, E.N., 1984. A reinterpretation of the banking crisis of 1930. Journal of Economic History 44, 119–38.

11

Liquidity, Efficiency, and Bank Bailouts

GARY B. GORTON AND LIXIN HUANG*

In the early 1980's high interest rates caused many U.S. savings and loan institutions to become economically distressed. At the height of the crisis, the period 1988–1992, an average of one bank or S&L was closed every day (Mary L. Bean et al., 1998).[1] But, what was to become the thrift crisis in the late 1980's has, in large part, been attributed to the fact that insolvent institutions were allowed to remain open, mostly due to the depleted resources of the Federal Savings and Loan Insurance Corporation. Insolvent thrifts were not promptly closed and their assets sold to new investors. The policy of allowing insolvent institutions to remain open was labeled a policy of "forbearance" and Edward J. Kane deemed the insolvent thrifts "zombies."[2] Eventually, the Resolution Trust Corporation

* We thank Franklin Allen, Jack Kareken, Richard Kihlstrom, Adriano Rampini, three anonymous referees, and seminar participants at the University of Minnesota and the Federal Reserve Bank of New York for comments and suggestions. Huang acknowledges partial research funding provided by a grant from City University of Hong Kong.

1. During the period 1980–1994, 1,617 banks with $302.6 billion in assets were closed or received assistance from the Federal Deposit Insurance Corporation. During the same period, 1,295 savings and loans, with $621 billion in assets, were closed by the Federal Savings and Loan Insurance Corporation (FSLIC) or the Resolution Trust Corporation, or received assistance from FSLIC; see Bean et al. (1998). On the thrift crisis generally, see R. Dan Brumbaugh, Jr. (1988), Edward Kane (1989), James Barth (1991), and Lawrence White (1991).

2. The term "zombie thrifts" became widely used and is now applied to similar banking situations. In a private communication, Ed Kane recalls having first used this term in a speech to the American Bar Association in 1986. It first appeared in published work in Kane (1987), a paper that was presented at the Western Economics Association in 1986.

was established to liquidate the assets of insolvent thrifts.[3] The bailout of the thrift industry ultimately cost $180 billion (3.2 percent of GDP); see Gerard Caprio and Daniela Klingebiel (1996).[4]

Prolonged and expensive government bailouts of financial intermediaries following banking crises have recently proliferated around the world, and it is not only transitional and emerging economies that have had such experiences.[5] As in the U.S. thrift crisis, the resolution of these crises typically involves the use of public money to subsidize the restructuring or disposal of impaired loans, a "bailout." In a survey of 120 banks in 24 developed countries in the 1980's and 1990's, Charles A. E. Goodhart (1995) found that two out of three failed banks were bailed out. These bailouts are expensive. In a sample of 40 such episodes, Patrick Honohan and Klingebiel (2000) found that, on average, countries spend 12.8 percent of their GDP cleaning up their banking systems. Stijn Claessens et al. (1999) set the costs at 15–50 percent of GNP. To emphasize, even developed economies other than the United States have faced large costs of bank bailouts. For example, Spain is estimated to have spent 16.8 percent of GNP on bailouts; Sweden, 6.4 percent of GDP; Finland, 8 percent of GDP. See Caprio and Klingebiel (1996).

In bank bailouts the government directly aids banks by buying equity, extending long-term loan guarantees to the banks, or buying bank loans at favorable prices. [Usually, nonperforming loans are purchased by the government at face value (see Daniel, 1997).] Sometimes government bonds are exchanged for bad bank loans. Often a public centralized asset management company is set up that uses government funds to lend to troubled banks against specific loan collateral or that buys the loans from the banks (e.g., see John Hawkins and Philip Turner, 1999; David Woo, 2000). For example, in the Asian crisis, government-owned asset management companies in Indonesia, Malaysia, Korea, and Thailand had,

3. The 1989 Financial Institutions Recovery, Reform and Enforcement Act (FIRREA) substantially changed the regulatory structure of the thrift industry. The Resolution Trust Corporation (RTC) was part of the 1989 law. The RTC was the government vehicle for selling the assets of closed thrifts.

4. According to Barth and Bartholomew (1992), as of 1992: "More than 500 institutions were closed at an estimated present-value cost in excess of $50 billion. Still another 500 or more institutions were open but insolvent at the end of the decade. These remaining candidates for closure will cost an estimated $100 billion or more . . ." (p. 37). Other estimates are considerably higher. For example, the *Wall Street Journal*, April 6, 1990, cites a Congressional Budget Office and General Accounting Office projection of a cost of $300 to $350 billion. See Kane and Min-Teh Yu (1996) for market value estimates.

5. Caprio and Klingebiel (1999) count 112 episodes of systemic banking crises in 93 countries since the late 1970's. Since 1980, at least two-thirds of International Monetary Fund member countries have experienced problems with their banking systems (see James A. Daniel, 1997). Also, see Carl-Johan Lindgren et al. (1999).

by April 1999, taken over bank assets with face values equivalent to 20, 17, 10, and 17.5 percent of the GDP of these respective countries (Lindgren et al., 1999).[6] An early, and influential, example of such a vehicle was the Reconstruction Finance Corporation (RFC) that President Herbert Hoover initiated during the onset of the Great Depression in the United States (see James Stuart Olson, 1977; Joseph Mason, 2001). The RFC is the (implicit or explicit) model for a large number of such vehicles in countries around the world, including the Resolution Trust Corporation, founded to use federal money to buy and then sell assets of insolvent U.S. savings and loans institutions (see Walker Todd, 1992). As we discuss later, government bank restructuring agencies are now commonplace.[7]

Why do government bailouts occur? Why does the government engage in forbearance, rather than simply closing insolvent banks and selling their assets to private investors immediately?[8] Are government bailouts efficient? The basic idea developed in this paper is that it is costly for private agents to be prepared to purchase substantial amounts of assets on short notice, such as the assets of the banking system (or a large part of the banking system), because liquidity is socially costly. Simply put, the sheer volume of the assets that need to be sold can be too large for private agents to absorb quickly. The resources of private agents are "illiquid." In order to make this point, we first must address the issue of "liquidity." What is "liquidity" and where does it come from? And, how does "liquidity" relate to "market efficiency"? The first part of the paper addresses these questions. The second part of the paper demonstrates how these answers are useful in understanding government bailouts. We show how the government can create liquidity and improve welfare. We go on to address the issue of whether

6. Hawkins and Turner (1999) list 15 countries that have recently restructured their banking systems and corporate sector with asset management companies; see their Table 12.

7. It is not only financial institutions that are bailed out. Nonfinancial firms are also sometimes bailed out directly. Nonfinancial firm bailout examples include firms that had outstanding commercial paper during the Penn Central crisis (see Charles Calomiris, 1994), the Chrysler Corporation (see Lee Iacocca and William Novak, 1986), and the airlines industry following the terrorist attacks of September 11, 2001. In these cases, the government provided loan guarantees to a syndicate of banks and, in the case of Penn Central crises, provided liquidity through the Federal Reserve's discount window lending. A related example is the case of the hedge fund Long Term Capital Management (LTCM), where the lenders in effect purchased the assets at the government's instigation (see Lowenstein, 2000). But, in most countries corporate restructuring is intimately related with bank restructuring because of the prevalence of bank loans; most of the nonperforming loans of a banking system are the obligations of nonfinancial firms that are no longer able to meet their debt payments (see Hawkins and Turner, 1999, and Mark Stone, 2002).

8. We do not address the issue of why banking systems come to be distressed. See Gorton and Andrew Winton (2003) for a survey of the literature on banking crises and banking panics.

the government has the resources to bail out the banking system. Perhaps a policy of "forbearance" is optimal. We provide conditions under which this is the case.

We start with a general setting in which agents who have invested in securities or projects sometimes need to sell them later. The price the projects fetch at that time depends upon the "liquidity" of the market. We present a general-equilibrium model in which not all assets can be used to purchase all other assets at every date. "Liquid" assets are the only readily available funds that can be used to purchase projects/securities from other agents. "Liquidity" then refers to the extent to which liquidity supply can meet liquidity demand. If insufficient liquid funds are available, then claims on projects cannot be subsequently transferred, because potential new owners have no way to buy them. Projects that cannot be transferred may decline in value. Recognizing that there may be opportunities at future dates to buy, some agents initially choose to invest in "liquid" assets. So, liquidity provision can be socially efficient.

There are two requirements for a model of liquidity. First, there must be a need to trade: at some dates there must be agents seeking to sell assets and there must be other agents who are willing to buy. Second, there is a restriction needed, namely, not all assets can be used to purchase all other assets at every date. Buyers must be restricted to making purchases only with certain assets, "liquid" assets. This restriction is akin to a cash-in-advance constraint. The constraint arises from the fact that other assets are not liquid; they cannot be used to trade at some dates. In the model here agents make investment choices at the initial date, choosing a long-term investment project or choosing a short-term investment project. The only purpose of selecting a short-term project is to get "liquid" assets to possibly buy a troubled long-term project at a later date. Long-term projects are financed by borrowing from lenders. Subsequently, long-term project borrowers learn whether their projects are high value or low value. Owners of low-value projects are "distressed" and may engage in moral hazard. But low-value projects may be recapitalized by being sold in the "liquidation" market to agents with available liquid capital. The only buyers available are agents who initially chose short-term projects for exactly this purpose ("vulture" investors).

It is costly for a society to always have liquid funds stand by in large amounts, as alternative investments (long-term projects) are (socially) more attractive, but "illiquid." The private provision of liquidity can be avoided, if, instead of projects being sold in the liquidation market, the original lenders are willing to forgive debt of borrowers with low-value projects. Such forgiveness dominates the private provision of liquidity since it does not involve inefficient "hoarding," i.e., investing in the short-term project to get liquid funds. But, forgiveness is not always in the interests of lenders. If forgiveness is not in their interests, then privately supplied liquidity is the only way to recapitalize

projects. Since this is socially costly, perhaps the government can improve welfare by supplying liquidity. We show what it means for the government to supply liquidity.

Owners of long-term projects that had a high value realization are privately illiquid because they cannot monetize their gains, i.e., their assets are illiquid. But, the government can monetize their gains by issuing securities backed by tax revenue collected from these agents later. The government securities are used to subsidize owners of low-value projects, while owners of high-value projects are taxed later. Low-value projects are recapitalized by the subsidies.

We then extend the analysis to consider possible systematic risk in the banking system and to investigate the phenomena of bank bailouts. As above, investment in projects at the initial date is financed through borrowing. Subsequently, long-term project owners learn the value of their projects and may desire to sell their projects in the liquidation market in order to obtain the recapitalization. But, the lending banks may also want to engage in moral hazard for the same reasons that their borrowers might engage in moral hazard. If banks suffer a negative shock to their capital (from other asset-side activities), then they may acquiesce in allowing their borrowers to engage in moral hazard rather than seeking to recapitalize them. In this case, banks are not interested in liquidating their borrowers' projects. Anticipating that this may occur affects the initial choice of investments, in particular the supply of liquidity, and can reduce social welfare. We provide conditions under which the government can improve welfare by bailing out banks.

There are many different notions of "liquidity" and "illiquidity" in the literature.[9] In the finance literature, liquidity is not explicitly modeled. Rather there are "noise traders" or "liquidity traders," modeled as exogenous random amounts of buy and sell orders. The other side of the market is the "market maker" who has an inventory that potentially can be long or short an infinite amount (see Albert S. Kyle, 1985). Because the market maker's inventory is infinite, the price set is equal to expected value of the payoff (conditional on available information). That is, the supply curve of liquidity is infinitely elastic so the price is not influenced by the size of the market maker's inventory. The model here obviously differs in important respects. Because the model is one of general equilibrium, the supply of liquidity will be determined endogenously and will not be

9. Douglas W. Diamond and Philip H. Dybvig (1983) view consumption smoothing as liquidity. Hugo A. Hopenhayn and Ingrid M. Werner (1996) develop a notion of liquidity based on search. In Gorton and George Pennacchi (1990) liquid assets are assets that minimize trading losses to uninformed traders when they trade in markets with privately informed traders. "Liquid assets" are not sensitive to private information because they are relatively riskless, like bank deposits. Andrea Eisfeldt (2002) presents a general equilibrium model of liquidity based on adverse selection. Agents are motivated to trade by changes in productivity and this interacts with desires to self-insure.

infinitely elastic. Because we explicitly model liquidity provision we can conduct welfare analysis.

Often, "liquidation" of projects is modeled in partial equilibrium as an exogenous value, interpreted as the value of the project in its next best alternative use. In reality, the next best alternative use is the value of the project to another agent, so the "liquidation value" of a project depends on the price that the project fetches in the market when it is sold to the other agent. And that price, in turn, depends on the supply of liquidity, that is, on the aggregate resources of those agents who can feasibly bid at that date. Andrei Shleifer and Robert Vishny (1992) observe that when a distressed firm needs to sell assets, the natural buyers are other firms in the same industry. The other firms may also be distressed, leading to the conclusion that the price of the assets being liquidated is partly a function of the available supply of liquidity, not just information about payoffs. We formalize this by showing how it can occur that liquidity is not in perfectly elastic supply; in that case, a project's "fundamental value" may be different from the price it trades for in the market, a price that depends upon the amount of liquidity available, or the market "depth." Thus, the notion of "market efficiency," i.e., the idea that prices are conditional expectations of project payoffs, requires perfectly liquid markets. Otherwise, markets are not price efficient, but there is no arbitrage possible because of the lack of liquid assets to conduct such a trade. We provide a model of this.

Intimately related to any notion of "liquidity" is an assumption of a shock that is the motivation for immediate selling or borrowing. Our focus on the transferability of distressed projects, or more specifically on the transferability of control rights to such projects, is quite different from the notion of liquidity in Bengt Holmström and Jean Tirole (1998) and Diamond and Raghuram G. Rajan (2001, 2002). In those papers, a "liquidity shock" is an event that causes firms to need to borrow extra funds from other firms or from consumers. This idea has its origin in Diamond and Dybvig (1983), although in that instance the shock was to consumers' consumption timing rather than to firms' investment opportunities. In our paper there is no investment boom-type shock (where firms suddenly need resources), but there is a shock to the value of assets in the hands of an entrepreneur for a given level of liabilities—a "capitalization shock." Then there is a potential need to sell the control rights to the project. Anticipating that this market may be open at the interim date, entrepreneurs choose at the first date whether to be either buyers or sellers in the secondary market. Thus, we endogenize the supply of liquidity.

The choice of shock, Diamond-Dybvig-type "liquidity shocks" versus "capitalization shocks" is important for policy results. For example, Diamond and Rajan (2001, 2002) also discuss recapitalization by the government, observing that it may be counterproductive because it increases the demand for liquidity without increasing the supply of liquidity. As a result, the interest rate rises and banks may

be forced into insolvency. Such distortionary government intervention should not occur in equilibrium. Our model is very different, partly because the shock is different. Unlike Diamond and Rajan, no firm is forced to liquidate; firms may choose to liquidate. Government intervention in our model is not distortionary because the price in the liquidation market adjusts so that good projects will never be liquidated. In addition, in their model government intervention only affects the market for liquidity, but not the investment choice. In our model, in contrast, government intervention corrects investment inefficiency through its impact on the liquidation price. While ultimately an empirical matter, our view is that "capitalization shocks" are more important.[10]

Also, the notion of liquidity we develop is intuitive. It refers to the amount of resources standing ready to purchase the claims on projects should there be a desire to sell the projects at some date between initiation and final payoff. The motivation for projects to be sold is financial distress. As in Sanford J. Grossman (1988), some agents must commit at the initial date to have certain resources available at an interim date, should opportunities arise. We provide a model of this "liquidity-in-advance" constraint. Our focus enables us to extend the analysis to banking crises and bank bailouts. Since the model is one of general equilibrium, we can conduct welfare analysis.

The paper proceeds as follows. In Section 11.1 the model is presented and the assumptions discussed. In Section 11.2 we show that there is no need for liquidity, in the sense of agents standing ready to buy projects, if lenders are willing to forgive the debt of distressed borrowers. Debt forgiveness is a kind of liquidity provision on the part of the lender, and it can implement the first-best allocation. If debt forgiveness is not optimal for lenders, then there is a need for private agents to provide liquidity. This is studied in Section 11.3. The equilibrium in which liquidity is provided privately is analyzed. This allocation is not first best, but the government via bailouts may be able to improve welfare and achieve first best. A system of taxing and subsidizing amounts to liquidity creation, and we provide a condition under which first best can be achieved. It is studied in Section 11.4. Section 11.5 considers the extended case where bank lenders themselves may seek to engage in moral hazard, by acquiescing in allowing their distressed borrowers to engage in moral hazard. This results in an even less efficient outcome. However, under certain conditions, a government bailout can achieve first best. Section 11.6 concludes.

10. There are many other papers that make use of Diamond-Dybvig-type shocks, e.g., Sudipto Bhattacharya and Douglas Gale (1987), Diamond (1997), and Franklin Allen and Gale (1998), among others. These papers examine the consequences of an aggregate shortage of liquidity. Not only is our shock different, but also in our model the liquidity problem is generated by incentives to add risk when capital is inefficient. The moral hazard not only generates a demand for liquidity, but also restricts the supply of liquid.

11.1. THE MODEL

In this section we present the model and discuss the important assumptions.

There is a continuum of depositors, a continuum of entrepreneurs, and a continuum of banks in the economy. There are three dates: 0, 1, and 2. At date 0 entrepreneurs must choose one of two possible investment projects, either a long-term project or a short-term project. The long-term project requires an investment of \$1 at date 0. The short-term project only needs the entrepreneurs' human capital. The cash flow of the long-term project is realized at date 2. It can be high (H) with probability π or low (L) with probability $1 - \pi$. Information about the date 2 cash flow arrives at date 1. The cash flow of the short-term project is realized at date 1, which is a random variable equal to r, where r is uniformly distributed on the interval $[0, R]$. For simplicity, there is no available short-term project between dates 1 and 2.

11.1.1. Assumptions

We make the following assumptions on cash flows from the projects, operation of the projects, and how the projects can be financed.

Assumption 1. $\pi H + (1 - \pi)L > 1 + R/2$. *That is, the long-term project has a higher expected cash flow than the short-term project.*

Because the long-term project offers superior returns, we refer to investing in the short-term project as "hoarding." As we will see, the only reason any entrepreneur will invest in the short-term project is that it enables him to purchase a long-term project in the low state at date 1.

Assumption 2. *Each entrepreneur can only manage one project at a time. Projects are not divisible.*

Assumption 3. *Neither the date* 1 *states* (H *or* L), *nor the cash flows realized at date* 2, *are contractible. Therefore, outside equity financing is not feasible.*

Entrepreneurs have no resources at date 0, so they need to borrow from banks. The credit market is competitive at date 0. Therefore, entrepreneurs receive the entire expected surplus and lenders earn zero expected profits. For simplicity, all agents are risk neutral and the riskless interest rate is zero.

At date 0 investment decisions are taken and entrepreneurs who choose long-term projects borrow from the lenders. Because states and cash flows are not verifiable, we only allow debt contracts. Gorton and James Kahn (2000) rationalize the use of a debt contract in this setting, where there is no interim cash flow from the project, but where there is a moral hazard problem. They show the optimality of bank debt, where the debt can be renegotiated at the interim

date. Initial terms are not set to price default risk but rather are set to efficiently balance bargaining power in later renegotiation, and renegotiated interest rates may not be monotonic in firm risk.

At date 1 entrepreneurs operating short-term projects receive the realization of the projects' cash flows. Entrepreneurs operating the long-term project receive no cash flows at date 1, but they learn the realization of the state, H or L. This information is also observable by the lenders, but by Assumption 3 it is not contractible. Other agents, in particular the government, do not observe the realized project states at date 1.

Denote by α the fraction of entrepreneurs that take the long-term project at date 0; the fraction of the entrepreneurs that take the short-term project at date 0 is $1-\alpha$. After each entrepreneur chooses a project type, those choosing the long-term project sign a debt contract with one of the lenders. The face value of the debt contract, F, will be determined in equilibrium.

There is a potential moral hazard problem because the entrepreneurs may engage in asset substitution at date 1. Each entrepreneur, regardless of which project has been selected at date 0, has access to a constant returns to scale risk-adding technology. By adding risk, one unit of certain value generates either a very large value, T, with probability δ or a value of zero with probability $1-\delta$.

Assumption 4. $\delta T < 1$. *That is, adding risk to the project is inefficient.* $1-\delta T$ *is the expected loss per unit from asset substitution.*

Suppose an entrepreneur owes an amount f and he learns that his cash flow at date 2 will be v. If he does not add risk, his payoff is: $\max[v-f,\ 0]$. If he adds risk, his expected payoff is: $\max[\delta(vT-f),\ 0]$. Suppose vT is greater than f. The entrepreneur is not going to add risk if and only if $\delta(vT-f) \leq v-f$. That is, the face value of the debt, f, cannot exceed γv, where $\gamma \equiv \frac{1-\delta T}{1-\delta}$.

Assumption 5. $\gamma L < 1$. *Since the debt face value is at least \$1, this assumption says that in the low state, the cash flow to the long-term project is not large enough to prevent entrepreneurs from adding risk. Thus, there is a moral hazard problem at date* 1.

Although entrepreneurs may have incentives to engage in moral hazard at date 1, *ex ante* they want to prevent it because they get the entire expected surplus and bear the entire potential loss caused by risk-adding. There are two possible ways to prevent the moral hazard problem. First, the contract between the lender and an entrepreneur can be renegotiated at date 1 to remove the entrepreneur's incentive to engage in asset substitution. In particular, the lender can forgive some of the debt, by lowering the face value of the (pure discount) debt, and thus increase the entrepreneur's equity to eliminate his incentive to add risk. But this depends on whether the lender is willing to forgive debt. It can happen that the lender finds forgiveness unprofitable and renegotiation breaks down.

Secondly, the project may be "liquidated." "Liquidation" means that the "troubled" project, i.e., the project of an entrepreneur who will otherwise add risk, is sold to a new owner at date 1. It is better to sell the troubled project if the liquidation value is higher than the continuation value of the project with risk being added. On the other hand, selling the project to a new owner means that the assets are redeployed, but does not necessarily mean that the new owner will add no risk. If the new owner has to borrow at date 1 to buy the project, then the lender will have to ensure that the new owner has enough equity in the project so that they do not have an incentive to add risk upon buying the project.

A central question is: Who will buy troubled projects at date 1? Because each entrepreneur can only manage one project at a time, buyers of projects at date 1 can only be entrepreneurs who undertook the short-term project at date 0. We will call them "liquidity suppliers" since it is the availability of their resources at date 1 that can allow project ownership to be transferred. Clearly, the price at which a troubled project can be sold at date 1 will depend on the demand and supply of liquidity and this price will be determined in equilibrium. In addition to the assumptions we have made, we make the following assumptions about the cash flows from the long- and short-term projects to make our model more interesting.

Assumption 6. $(\pi + \delta(1-\pi))\gamma H \geq 1$. *Due to the possible moral hazard problem, the face value of the debt cannot exceed γH. Otherwise entrepreneurs will add risk even in the high state. This assumption guarantees that the cash flow in the high state is large enough that the bank is willing to lend at date* 1, *even though entrepreneurs might engage in risk-adding when they are in the low state.*

Assumption 7. $TL > \gamma H$. *That is, if an entrepreneur engaging in moral hazard is lucky, there is a chance that risk-adding can produce a cash flow high enough that his equity value is positive after repaying the face value of the debt.*

Assumption 8. $\delta T > \gamma$. *This assumption is equivalent to $\delta TL > \gamma L$. δTL is the expected date* 1 *value of a long-term project in the low state when risk has been added. γL is the maximum of the face value that a bank can charge (possibly after forgiveness) such that the entrepreneur has no incentive to add risk. By assuming $\delta TL > \gamma L$, we make it possible that forgiveness is not in the interests of the lenders.*

Assumption 9. $R > L - \gamma TL$. *R is the maximum payoff from the short-term project. This assumption implies that if there is an excess supply of liquidity, the liquidation price can be as large as the "fundamental" value of the asset. This will become clear when we analyze the liquidation market in Section 11.3.*

To summarize, the sequence of events at date 1 is as follows.

(i) An entrepreneur who invested in the short-term project at date 0 receives a cash flow of r at date 1.
(ii) News arrives about whether the date 2 cash flow of each long-term project will be high (H) or low (L).
(iii) Owners of troubled long-term projects renegotiate with their lenders to try to reduce their debt burden.
(iv) Depending on the outcome of the renegotiation, an entrepreneur may "liquidate" his project, i.e., sell it in the market, or he may continue until date 2 (possibly adding risk) to receive the final cash flow. If the project is sold in the liquidation market, then the new owner operates the project (possibly adding risk), and receives the final cash flow at date 2.
(v) If the project is liquidated, then a final payment (that was previously renegotiated) is made (at date 1 or possibly at date 2) to the lender to settle the outstanding loan. The new owner then operates the project.

At date 2, the cash flows of the long-term projects are realized and lenders are repaid.

By Assumption 1, the first-best outcome would be for all the entrepreneurs to invest in the long-term project. But, because of the moral hazard problem, this may not be the outcome of private decisions. Some entrepreneurs may choose to invest in the short-term project, in order to act as liquidity providers at date 1. The date 1 opportunity may be sufficiently valuable to make "hoarding" at date 0 attractive.

11.1.2. Discussion of Assumptions

Assumption 2 (projects are indivisible and entrepreneurs can only run one project at a time) means that markets are incomplete. Note that some entrepreneurs who chose the long-term project at date 0 receive good news at date 1 so they have projects worth H. In other words, at date 1 they have a capital gain relative to the expected value of the project as of date 0. If their equity is high enough, they could issue subordinated debt to buy another troubled project. But, since projects are indivisible, their equity value has to be high enough to buy an entire firm. Whether this can happen or not depends on parameter values. For simplicity, we just assume that their limited human capital only allows them to manage one project at a time.[11] Alternatively, the long-term projects of

11. Alternatively, one can imagine more complicated explanations for why entrepreneurs who have realized H cannot credibly issue securities to buy projects at date 1, e.g., asymmetric information could be introduced.

different borrowers can be interpreted to be different (say, different industries), and entrepreneurs in the high state in one industry do not have the expertise to manage the projects in the low state, implicitly in another industry.[12]

There are several interpretations of the random cash flow return to the short-term project, r, realized at date 1. A straightforward interpretation is that the short-term projects literally are a technology with some uncertainty. Alternatively, one can think of the short-term projects as producing R, but entrepreneurs have a random consumption need or production cost uniformly distributed in $[0, R]$ which is subtracted from R, leaving r. As a practical matter we have in mind agents who hoard cash, government securities, short-term commercial paper, and so on—securities with low yields—in order to be able to buy distressed projects. We assume that the new owner generates the same cash flow. Cases of higher or lower cash flow can easily be solved with slight modification.

A model of liquidity needs a motive for some agents to sell assets. Here moral hazard provides the motivation for selling projects at date 1. The "capitalization shock" generates liquidity demand from the asset side, which differs from liability side shocks, such as random preferences over the timing of consumption. The specific motivation of moral hazard is important in Section 11.5, when we consider the banking system.

Another issue concerns the shocks to the long-term projects. They are idiosyncratic and independent of the random payoffs to the short-term projects. However, we could have systematic shocks and idiosyncratic shocks, plus correlation between the cash flows from the long-term and short-term projects without changing the main results. For simplicity, we do not include systematic shocks and correlation in the model.

11.2. DEBT FORGIVENESS

To avoid the moral hazard problem, borrowers need an equity injection at date 1 if their projects are in the low state. If the project is sold in the liquidation market, then a new owner injects equity. But, the lender is an alternative source of equity, in the sense that debt forgiveness creates equity for the original owner. To begin the analysis, we analyze the case where equity is injected via the lender forgiving some of the debt. In this case, all entrepreneurs will invest in the long-term project at date 0, which is the first-best outcome, as there is no need for private liquidity provision at date 1. Investment in the short-term project is dominated.

12. Holmström and Tirole (1998) have a similar assumption. Lucky firms (with additional liquidity) are not allowed to take over unlucky firms (short of liquidity). In Diamond and Rajan (2002) banks find out at an interim date whether their loan maturities are long or short; banks cannot insure against the risk of having a long maturity portfolio realization at this date.

We start by solving for the subgame equilibrium at date 1. At date 1, if a long-term project is in the high state, then nothing happens (by Assumption 6, there exists a face value of the debt, F, such that an entrepreneur has no incentive to add risk in the high state.) However, if a long-term project is in the low state, L, then the project is worth L if no risk as added and is worth δTL if risk is added. By Assumption 5 borrowers will add risk if they retain ownership of the project without any new equity.

Renegotiation between the borrower and the lender has three possible outcomes. First, the original owner may continue until date 2 without adding risk if the lender forgives some debt. Second, the project may be sold or liquidated, in which case the original owner receives the price Q (the liquidation value). However, this entrepreneur owes the lender F, which must be paid at date 1 if the debt is short term or at date 2 if the debt is long term. Renegotiation must allocate Q between the borrower and the lender. Finally, the project may remain in the original owner's hands with risk being added. Equilibrium at date 1 involves determining which of these outcomes is the result of renegotiation between borrowers and lenders.

In order for the first possibility to occur, the lender must forgive part of the debt. This is an equity injection, a subsidy granted to the entrepreneur by the lender, to induce the entrepreneur not to add risk. By the analysis above (just before Assumption 5), the lender must agree to lower the face value of the debt to γL to remove the entrepreneur's incentive to add risk. The lender is willing to forgive debt, that is, reduce the face value from F to γL, if and only if γL is greater than δF, the expected payoff to the lender when risk will be added. The following lemma provides the condition under which debt forgiveness is feasible.

LEMMA 1. *Debt forgiveness is feasible if and only if* $\gamma L \geq \frac{\delta}{\pi + \delta(1-\pi)}$.

Proof. See Appendix.

If debt forgiveness is feasible, then there will be no need for liquidity provision at date 1.[13] Since, *ex ante*, the long-term project is more efficient than the short-term project, no entrepreneur will choose the short-term project at date 0; all entrepreneurs will choose the long-term project. Lenders only receive the face value of the debt, F, in the high state, which happens with probability π. In order to make lenders break even, F has to satisfy the following condition: $\pi F + (1 - \pi)\gamma L \geq 1$, or $F \geq \frac{1-(1-\pi)\gamma L}{\pi}$. Proposition 1 summarizes the above analysis:

13. Later we will show that the liquidation price cannot exceed L. Therefore, the entrepreneurs who choose the long-term project at date 0 also choose debt contracts under which renegotiation results in forgiveness of debt.

PROPOSITION 1. *If* $\gamma L \geq \frac{\delta}{\pi + \delta(1-\pi)}$, *then all entrepreneurs choose the long-term project at date 0. The face value of the debt is* $\frac{1-(1-\pi)\gamma L}{\pi}$. *This is the first-best outcome.*

The proposition says that as long as debt forgiveness at date 1 is feasible, it will occur. The details of the division of the surplus between the lender and the borrower do not matter. Whatever the division, based on relative bargaining power, the equilibrium will be the same, as the entrepreneurs price this in getting the entire expected surplus at date 0.

Debt forgiveness by the lender is a kind of liquidity provision because the lender is essentially refinancing the project, taking into account the new information, namely the state L, and the problem of moral hazard.[14] Liquidity provision is an important function of banks.[15] If forgiveness by the lender is feasible, then there is no need for a secondary market to refinance the project (by selling it to another party). Proposition 1 allows us to identify the condition under which the provision of liquidity (by agents other than the lenders, i.e., by entrepreneurs who hoard liquid assets) is socially valuable. According to the proposition, debt forgiveness solves the moral hazard problem, and there is no need for liquidity provision, so long as $\gamma L \geq \frac{\delta}{\pi + \delta(1-\pi)}$. If this is not the case, then there is a need for the liquidation market. *Ex ante*, the owners of long-term projects get a surplus if the liquidation price prevailing in the secondary market is higher than the value of project with risk being added. The liquidity suppliers gain a profit if the liquidation price is less than the value of the project free from added risk. At date 0 entrepreneurs make project choices and at the same time choose to be liquidity demanders or liquidity suppliers. At date 1, a secondary market will arise endogenously. This is analyzed in the next section with the assumption that $\gamma L \leq \frac{\delta}{\pi + \delta(1-\pi)}$, i.e., debt forgiveness is not feasible.

11.3. THE MARKET FOR LIQUIDITY

If lenders are unwilling to forgive debt, then equity will have to be created in some other way as a solution to the moral hazard problem. The alternative is to sell the project to another, better capitalized, entrepreneur at date 1. We refer to this secondary market as the "liquidation market." In other words, at date 1 there

14. The case of the hedge fund Long Term Capital Management (LTCM) is a recent example of forgiveness by lenders (see Roger Lowenstein, 2000).

15. The empirical results of Scott Lummer and John McConnell (1989) suggest the positive announcement effect associated with bank loans is due to loan renewals, rather than the initial loan, consistent with bank debt forgiveness being important. See Gorton and Winton (2003) for a survey of the related literature.

is a market in which owners of long-term projects can sell their projects to other entrepreneurs with available resources to purchase the project.

11.3.1. Preliminaries

Because we are assuming that debt forgiveness is not feasible, the project can either be liquidated at price Q, or can be continued with risk being added, in which case it has an expected cash flow of δTL. Actually, whether the project will be liquidated only depends on the liquidation price Q. It does not depend on the maturity of debt contracts and the assignment of bargaining power when renegotiation occurs at date 1. So long as Q is greater than δTL, the lender and the borrower can always reach an agreement to liquidate the project and split the surplus. For simplicity, we assume that the debt contracts are long-term and the borrower has all the bargaining power. When renegotiation occurs at date 1, he makes a take-it-or-leave-it offer to the lender. The lender gets δF (the lender's expected payoff if the project continues with risk being added) and the borrower gets $Q - \delta F$.

11.3.2. The Liquidation Market and Liquidation Prices

Potential liquidity suppliers at date 1 are those entrepreneurs who invested in the short-term project at date 0. At date 1, each of them has realized a cash flow of r, drawn from a uniform distribution on $[0, R]$. If r is small, then the entrepreneur will not have enough to afford the liquidation price. Therefore, some of the entrepreneurs may have to borrow in order to buy a troubled project. Buyers themselves also face the moral hazard problem since they too have access to the risk-adding technology. Although the realized r is not publicly observable, we assume that if a project buyer borrows from a lender to buy a project, how the loan is used is verifiable. In other words, the amount borrowed can only be used to buy the project. In this way, if a buyer borrows B and buys a project at price Q, the lender knows the borrower's realized r is at least $Q - B$. Therefore, it can be determined whether potential buyers have incentives to add risk to the project and whether the loan to the buyer is safe.

A buyer has no incentive to add risk if and only if the face value of the debt B is small enough such that $L - B \geq \delta(TL - B)$, i.e., $B \leq \gamma L$. Therefore, the buyers who have a realized $r \geq Q - \gamma L$ are not going to engage in moral hazard and loans to them are safe. Other buyers do not have enough equity and have incentives to add risk once they get the control of the projects.

Because of the buyers' potential moral hazard problems, not every entrepreneur who hoarded liquid assets can be a liquidity supplier. Liquidity supply at date 1 depends on the buyers' ability to buy. "Ability" means how much

equity they have, that is, the size of the realized return from their short-term project. The next lemma characterizes their ability to supply liquidity at a given price.

LEMMA 2. *Suppose the liquidation price of a project is Q. Then:*

(*i*) *If* $Q > L$, *then there will be no liquidity supply in the liquidation market.*
(*ii*) *If* $\delta TL < Q \leq L$, *then only those buyers with realized cash flows* $r \geq Q - \gamma L$ *are able to supply liquidity.*
(*iii*) *If* $Q < \delta TL$, *then all buyers can be liquidity suppliers.*

Proof. See Appendix.

We know that a troubled project is liquidated if and only if the liquidation price Q is greater than or equal to δTL. So, the liquidity demand curve (or the project supply curve) is perfectly elastic at price δTL. By Lemma 2, we know that no entrepreneur is willing to pay more than L to buy a project. Moreover, as the liquidation price declines from L to δTL, more and more entrepreneurs are willing to buy the projects. Thus, there is a downward-sloped liquidity supply curve (or a downward-sloped project demand curve). Combining liquidity demand with liquidity supply determines the liquidation price in the secondary market, which depends on the fraction of entrepreneurs taking the long-term project at date 0.

LEMMA 3. *At date* 1, *the price in the liquidation market, Q, will be:*

$$Q = \delta TL$$

$$\text{if } (1-\alpha)\left(1 - \frac{\delta TL - \gamma L}{R}\right) \leq \alpha(1-\pi)$$

$$Q = \gamma L + R\left(1 - \frac{\alpha(1-\pi)}{1-\alpha}\right)$$

$$\text{if } (1-\alpha)\left(1 - \frac{L - \gamma L}{R}\right)$$

$$< \alpha(1-\pi) < (1-\alpha)\left(1 - \frac{\delta LT - \gamma L}{R}\right)$$

$$Q = L$$

$$\text{if } \alpha(1-\pi) \leq (1-\alpha)\left(1 - \frac{L - \gamma L}{R}\right).$$

Proof. See Appendix.

We define the "liquidity discount" to be the difference between the "fundamental" value of the project, namely L, and the liquidation price Q. Lemma 3

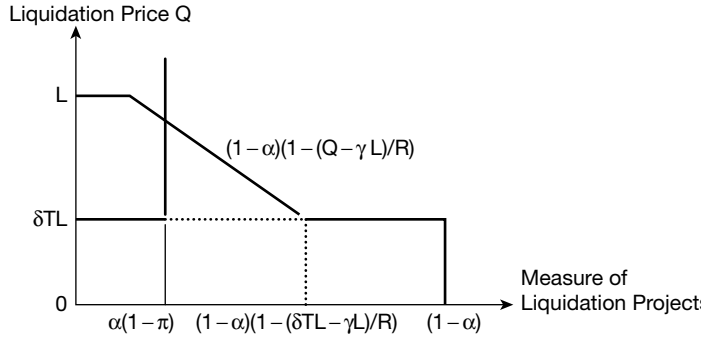


Figure 11.1 Liquidation Market Equilibrium.
NOTES: When α is in the intermediate range, liquidity demand curve intersects liquidity supply curve when it is still sloped down. The liquidation price is between δTL and L; the liquidity premium is between zero and $L - \delta TL$.

shows that a liquidity discount can arise in equilibrium. That is, the equilibrium price in the liquidation market, Q, can be *below* the value of the project if risk is not added, L. This discount is necessary to entice liquidity suppliers to invest in the short-term project in order to buy the projects.

To emphasize that the market price of the equity sold in the liquidation market has a price that depends on liquidity, we further examine the relationship between the price and the supply of liquidity with a figure. Figure 11.1 shows the secondary market equilibrium with respect to different levels of α, the fraction of entrepreneurs taking the long-term project. (So, $1 - \alpha$ is the fraction that has chosen short-term projects in order to later be liquidity suppliers.) When α is very small, liquidity demand is small while liquidity supply is large, and the liquidation price is at its maximum L. In this case, the liquidity discount is equal to zero (the vertical part of the liquidity demand curve intersects the supply curve at L). When α is very large, liquidity demand is large while liquidity supply is small, and the liquidation price is at its minimum δTL (the vertical part of the liquidity demand curve shifts right and intersects the supply curve at δTL). The liquidity discount is at its maximum $L - \delta TL = L(1 - \delta T)$. When α is in the medium range, there is an interior equilibrium liquidation price, Q, at which all the liquidity demand can be satisfied while liquidity supply is downward sloped. In this case the liquidity discount is between zero and $L(1 - \delta T)$, as shown in the figure.

In the liquidation market, ownership claims, i.e., equity claims, to the project are sold and a new entrepreneur acquires control rights. The new entrepreneur has the right to choose whether to add risk or not. In equilibrium no risk is added. In the finance literature, when "market efficiency" is mentioned, the transfer of control rights, which usually results in a change of the asset value, is typically not considered. Moreover, the supply of liquidity is perfectly elastic. Here, due to

the moral hazard problem and the limited liquidity supply, the notion of "market efficiency" is altered. Although, in equilibrium, risk will not be added and the continuation value of the project is L, there is a liquidity discount. *Ex post*, the liquidity discount reflects insufficient liquid assets in the market. *Ex ante*, the liquidity discount is necessary to compensate the liquidity suppliers because there is a cost associated with supplying liquidity. We next solve the entrepreneur's date 0 decision problem.

11.3.3. Initial Investment Choices

At date 0, entrepreneurs have rational expectations about how the liquidation price is formed in the secondary or liquidation market of date 1. An entrepreneur makes his project choice, taking other entrepreneurs' choices as given. We solve for the date 0 equilibrium in the following proposition (under the maintained assumption that debt forgiveness is not feasible at date 1).

PROPOSITION 2. *Suppose that* $\gamma L < \frac{\delta}{\pi+\delta(1-\pi)}$, i.e., *debt forgiveness is not feasible at date* 1. *Then:*

(*i*) *If* $\pi H + (1-\pi)\delta TL > 1 + \frac{R}{2} + (L - \delta TL) \times \left(1 - \frac{\delta TL - \gamma L}{R}\right)$, *then all the entrepreneurs choose the long-term project at date* 0. *In this case, there will be no liquidity supplied at date* 1, *and risk will be added to all projects realizing the low state at date* 1.

(*ii*) *If* $\pi H + (1-\pi)\delta TL \leq 1 + \frac{R}{2} + (L - \delta TL) \times \left(1 - \frac{\delta TL - \gamma L}{R}\right)$, *then there is a fraction* $\alpha^* \in (0,1)$ *of entrepreneurs who choose the long-term project, and a fraction* $(1 - \alpha^*)$ *of entrepreneurs who choose the short-term project at date* 0. $\alpha^* = \frac{R+\gamma L - Q^*}{(2-\pi)R+\gamma L - Q^*}$, *where,* $Q^* \in [\delta TL, L]$, *defined in the Appendix, is the liquidation price at date* 1.

Proof. See Appendix.

Proposition 2 relates the primitives of the economy to the incidence of liquidation. Entrepreneurs compare the expected payoffs on the two projects available to them at date 0, taking into account the possible transactions in the liquidation market. The short-term project has a return of r over the first period and a return over the second period that depends on the profitability of buying a distressed project at date 1 and operating it until date 2. If these returns are too low, compared to the long-term project (when risk is added in case L is realized), then all entrepreneurs will still take the long-term project. Otherwise there is an active secondary market and some entrepreneurs engage in each activity. In that case, supplying liquidity is both privately profitable and socially efficient.

11.3.4. Comparative Statics

Proposition 2 provides the solution to the initial general-equilibrium investment problem. In equilibrium, entrepreneurs are indifferent between investing in the long-term project or in the short-term project at date 0; they get the same expected payoff *ex ante.* By assumption, entrepreneurs get the entire expected surplus, thus their expected payoff is a measure of social welfare. The expected payoff to entrepreneurs who take the long-term project is $\pi H + (1 - \pi)Q^*$, and the expected payoff to entrepreneurs who take the short-term project is $1 + \frac{R}{2} + (L - Q^*)\left(1 - \frac{Q^* - \gamma L}{R}\right)$. Social welfare, the liquidation price, Q^*, and thus the liquidity discount, $L - Q^*$, all depend on the parameter values: π, R, and δ. π measures the quality of the long-term project; R measures the quality of the short-term project; and δ measures the severity of the moral hazard problem. Proposition 3 and Corollary 1 show the comparative statics:

PROPOSITION 3. *Social welfare is increasing in π and R, and decreasing in δ.*

COROLLARY 1: *The liquidation price Q^* (the liquidity discount, $L - Q^*$) is increasing (decreasing) in R and decreasing (increasing) in π and δ.*

Proof. See Appendix.

$L - Q^*$ is the equilibrium liquidity discount, which is a measure of how profitable it is to hoard assets by investing in the short-term project. It is also the necessary compensation a liquidity provider requires to sacrifice the long-term project. The larger the difference between the expected payoff from a long-term project (in the absence of the moral hazard problem) and that from a short-term project, the higher the required liquidity discount. When δ increases, γ increases, and thus the threshold to be a liquidity supplier is increased. The liquidity discount has to be increased to compensate for the decrease in the probability of being a liquidity supplier (a lower Q^*). A larger π increases the expected payoff of the long-term project and decreases liquidity demand, thereby resulting in a higher liquidity discount (a lower Q^*). On the other hand, a higher R increases the expected return of the short-term project and also increases the potential to become a liquidity supplier; therefore it results in a lower liquidity discount (a higher Q^*).

The effects of R, π, and δ on welfare are quite intuitive. When R and π increases, the overall investment quality in the economy is improved. Hence, welfare increases. When δ increases, the reservation price δTL is higher and more short-term investors are needed to purchase the distressed long-term projects. More investment in dominated short-term projects results in lower welfare.

11.3.5. Summary

Entrepreneurs cannot buy insurance at date 0 against declines in the value of their equity at date 1. Nor can entrepreneurs with high-value projects at date 1 use that value to inject equity into the low-value projects. These markets do not exist and as a result of these missing markets, private liquidity provision can be efficient. The existence of a liquidity discount is a measure of the shadow price of the liquidity-in-advance constraint. To the extent that this constraint binds, the market price of project equity at date 1 reflects this constraint, not just the expected payoff on the project.

However, the government may be able to improve upon the private allocation because hoarding by investing in the short-term project is dominated by investment in the long-term project. We now analyze the role of the government.

11.4. GOVERNMENT BAILOUTS

Because of the moral hazard problem, the first-best outcome cannot be reached in the equilibrium studied above. The private supply of liquidity is inefficient since some investments are made in short-term projects, which are *ex ante* dominated by long-term projects. Can the government improve the efficiency of the economy? If the entrepreneurs had enough capital at date 0, then there would be no need for them to borrow, and the first-best outcome could be reached. Suppose the lenders are banks.[16] If the government has the power to tax bank depositors and subsidize the entrepreneurs at date 0 then they would have enough equity to avoid borrowing and the moral hazard problem would never arise. But, such transfers require that date 0 endowments be verifiable. In our model, entrepreneurs get the entire surplus. Depositors will not save their endowments in the bank if they anticipate the government is going to tax their savings. Therefore the subsidies have to be financed via taxing entrepreneurs in high states at the final date.

If the government can tax entrepreneurs with high returns and subsidize entrepreneurs with low returns, the government can at least partially improve efficiency by eliminating the incentives to add risk from some of the entrepreneurs in the low state. Unfortunately, the states of the long-term projects are only observable by banks and entrepreneurs and they are not verifiable. If the government cannot observe the states of projects at date 1, it has to design a screening mechanism to determine entrepreneurs are in the high state and which entrepreneurs are in the low state. We will show that the government

16. At this point, we assume lenders are banks, to foreshadow the analysis of bank bailouts in Section V.

can screen the banks by offering to buy the loans. In this section we examine such government bailouts.

11.4.1. Government Liquidity Provision

The government bailout mechanism works as follows. At date 1, the government offers to buy loans from banks. Each bank can either sell its loans to the government at a specified price P, or pay a tax, t, at date 2 if it does not sell its loans to the government at date 1. Once the government holds the loans, it can forgive a fraction of the liabilities of the troubled projects to remove the entrepreneurs' incentives to engage in moral hazard.

Alternatively, the government can offer subsidies to the banks (e.g., loan guarantees). To receive a subsidy, the bank has to lower the face value of its troubled loans. In equilibrium, only the banks with troubled loans will accept the offer from the government. In this way, the government can distinguish the high state projects from the low state projects and make transfers to improve efficiency. Whether government intervention can generate the first-best outcome depends on how much tax revenue it can collect from projects in the high state. The government certainly does not want the high-value projects to suffer from moral hazard problems and thus it must ensure that the owners of high-value projects pay less than γH at date 2 (i.e., the face value of the debt cannot exceed γH). If government intervention alone cannot generate the first-best outcome, there may still be a need for private liquidity to be supplied at date 1.

PROPOSITION 4. *If $\gamma(\pi H + (1-\pi)L) \geq 1$, then a government bailout can generate the first-best outcome as of date 0. At date 0, entrepreneurs and banks sign debt contracts with a face value of $F = \frac{1-(1-\pi)\gamma L}{\pi}$. At date 1, the government offers to buy the loan at price $P = \frac{\delta}{\pi+\delta(1-\pi)}$. Banks with troubled projects sell their loans to the government and banks with high state projects retain their loans and pay a tax of $t = \frac{1-(1-\pi)\gamma L}{\pi} - \frac{1}{\pi+\delta(1-\pi)}$ at date 2.*

Proof. See Appendix.

If $\gamma(\pi H + (1-\pi)L) < 1$, then government intervention cannot produce the first-best outcome because the government cannot levy enough taxes on the entrepreneurs in high states. Since the government cannot subsidize (or bail out) all the troubled projects at date 1, some of the troubled projects will suffer from the moral hazard problem if there is no private liquidation market. Whether there are entrepreneurs willing to take the short-term project and supply liquidity depends on the expected payoffs from the short-term project and the long-term project. We assume the government randomly chooses which projects to bail out when it does not have enough resources. The government determines the optimal bailout policy to maximize social welfare by choosing the loan price,

P, tax on high state projects, t, and the fraction, ω, of projects to bail out. So, the government's objective function is as follows:

$$\begin{aligned}
&\text{Max}_{\omega,P,t}\ \pi H + (1-\pi)(\omega L + (1-\omega)Q)\\
&s.t.\quad (1)\ F \le \gamma H\\
&\qquad (2)\ P \ge \delta(F - t)\\
&\qquad (3)\ \pi(F-t) + (1-\pi)\times(\omega P + (1-\omega)\delta(F-t)) = 1\\
&\qquad (4)\ \pi t \ge (1-\pi)\omega(P - \gamma L)\\
&\qquad (5)\ Q = \delta TL \text{ if } \pi H + (1-\pi)\times(\omega L + (1-\omega)\delta TL > 1\\
&\qquad\qquad + \frac{R}{2} + (L - \delta TL)\left(1 - \frac{\delta TL - \gamma L}{R}\right)\\
&\qquad (6)\ Q \text{ solves } \pi H + (1-\pi)\times(\omega L + (1-\omega)Q) = 1\\
&\qquad\quad + \frac{R}{2} + (L - Q)\left(1 - \frac{Q - \gamma L}{R}\right)\\
&\qquad\quad \text{if } \pi H + (1-\pi)\times(\omega L + (1-\omega)\delta TL) \le 1\\
&\qquad\quad + \frac{R}{2} + (L - \delta TL)\left(1 - \frac{\delta TL - \gamma L}{R}\right).
\end{aligned}$$

Constraint (1) says that F must be less than γH, to prevent entrepreneurs in the high state from engaging in moral hazard. Constraint (2) says that banks are willing to sell loans for no lower than $\delta(F - t)$, which is the amount they can receive if low state projects are liquidated or risk is added to them. Constraint (3) says that banks must break even. Constraint (4) is the government budget constraint: tax revenue collected from the measure of projects with the high value, πt, must be enough to cover the fraction of projects that the government chooses to subsidize. Each troubled project is subsidized by the amount $P - \gamma L$, and $(1 - \pi)\omega$ is the measure of low state projects subsidized. The final two constraints, (5) and (6), are participation constraints for private agents to supply liquidity. These are functions of the equilibrium price of projects at date 1, Q, which in turn depends upon the fraction of projects that the government subsidizes.

The following proposition characterizes the situations in which the liquidation market coexists with a government bailout at date 1.

PROPOSITION 5. *Suppose* $\gamma(\pi H + (1 - \pi)L) < 1$. *Define* $\omega^* \equiv \frac{\pi((\pi+(1-\pi)\delta)\gamma H - 1)}{(1-\pi)(\delta - (\pi+(1-\pi)\delta)\gamma L}$. *Then*

(*i*) *If* $\pi H + (1-\pi)[\omega^* L + (1-\omega^*)\delta TL] \ge 1 + \frac{R}{2} + (L - \delta TL)\left(1 - \frac{\delta TL - \gamma L}{R}\right)$, *then all the entrepreneurs choose the*

long-term project at date 0. *In this case, there will be no liquidity supplied at date* 1. *The government subsidizes a fraction* ω^* *of the troubled long-term projects and the remaining fraction* $1-\omega^*$ *of the troubled long-term projects will suffer from the moral hazard problem.*

(*ii*) *If* $\pi H + (1-\pi)[\omega^* L + (1-\omega^*)\delta TL] < 1 + \frac{R}{2} + (L-\delta TL)\left(1 - \frac{\delta TL - \gamma L}{R}\right)$, *then there is a fraction* α^* *of entrepreneurs choosing the long-term project, and a fraction* $(1-\alpha^*)$ *entrepreneurs choosing the short-term project at date* 0.

$$\alpha^* = \frac{R + \gamma L - Q^*}{(1 + (1-\pi)(1-\omega^*))R + \gamma L - Q^*},$$

where $Q^* \in [\delta TL,\ L]$, *defined in the Appendix, is the liquidation price at date* 1.

Proof. See Appendix.

Depending on the parameters, there may or may not be private liquidity provision. In either instance, not all entrepreneurs with troubled projects can be bailed out. The government cannot afford that, as there are not enough resources that can be taxed to subsidize the troubled entrepreneurs. The government randomly bails out the maximum number of troubled bank projects that it can afford, and in the remainder either risk is added or they go to the liquidation market. This may be viewed as a type of "forbearance," that is, the government policy results in some borrowers engaging in moral hazard. But, this is the socially optimal second-best policy.

11.4.2. Taxes and Government Bonds

The government can improve matters because it has the power to overcome the market incompleteness. Entrepreneurs who invested in long-term projects at date 0, and then saw their prospects brighten because they realized H on their project, have valuable resources, the capital gains on their project. But, they have no way to monetize these gains to create "liquidity" and provide equity to the troubled projects. The government, however, can monetize these gains by issuing government bonds, as follows. The government buys troubled projects from banks, paying with the newly created bonds. The government then forgives some of the debt of the entrepreneurs, to prevent them from engaging in moral hazard. The newly created bonds, held by the banks, are paid off at date 2 with the tax revenues raised from the lucky entrepreneurs who realized H.

Alternatively, the government can tax banks that make loans to long-term projects at date 0. The government then hoards the tax revenue and uses it to finance the bailouts at date 1. Since the burden has to be borne by the

entrepreneurs, banks will raise the face value of the debt to cover the tax they pay at date 0. This would generate the same outcome as the government issuing bonds and tax at date 2.

The fact that the government maximizes social welfare is also important for the argument. Suppose there is a private insurer who signs a contract with the bank and the entrepreneur at date 0. The contract specifies that the bank has the right to sell the loan to the insurer at (prespecified) price $P = \delta(F - t)$ at date 1 (a put option). However, contingent on the sale of the loan, the insurer must lower the face value of the debt to γL. If the bank does not sell the loan to the insurer at date 1, it has to pay the insurer t at date 2. The problem with this contract is that it is subject to collusion between the profit-maximizing insurer and the bank. At date 1 the insurer can bribe the bank not to force the insurer to buy the loan, according to the contract. The insurer gives the bank $t + \varepsilon$ (ε is a very small number) at date 1 if the bank agrees not to sell the loan; the bank accepts. Thus the insurer avoids losing $P - \gamma L$. Therefore no bank sells loans and if entrepreneurs anticipate the collusion, they will not purchase insurance at date 0. The government scheme works because the government is concerned with social welfare and does not want the entrepreneurs to add risk. The private insurer maximizes profits not social welfare.

11.4.3. Discussion

There is a long history to government bailouts of banking systems, either directly or via an asset management company, which is set up for the purpose of relieving banks of bad loans by buying loans at a price that implicitly subsidizes the banks. The Reconstruction Finance Corporation loans to railroads during the Great Depression, the examples of Chrysler, Penn Central Railroad, and the current (post-September 11) bailouts of the airlines in the United States are examples of government liquidity provision, as described above. Claessens et al. (1999) describe corporate-sector bailouts in Indonesia, Korea, Malaysia, and Thailand. In the aftermath of the Asian Crisis, Indonesia, Korea, Malaysia, and Thailand established centralized asset management companies. As a percent of GDP, the amounts of bank assets purchased by these asset management companies were: Indonesia, 20 percent; Korea, 10 percent; Malaysia, 17 percent; and Thailand, 17.5 percent. See Lindgren et al. (1999) for details of the bailouts resulting from the Asian Crisis. There is, however, a range of ways in which the bailouts are accomplished. Surveying the experiences of 24 countries in the 1980's and early 1990's, Claudia Dziobek and Ceyla Pazarbasioglu (1998) write:

> Removing nonperforming loans from the banks' balance sheets and transferring them to a separate recovery agency can be an effective way of

> addressing the banks' solvency problems. . . . Loan workouts can be done by a central organization, usually operated by the state, or by special loan collection agencies tied to individual banks, an approach Sweden used successfully in 1991. The survey results suggest that the institutional setting does not matter. Some countries, including Chile, the Philippines, and the transition countries, approached the loan workout indirectly by providing debt relief to borrowers by engaging simultaneously in the restructuring of borrowing enterprises themselves (p. 7).

(Also see Dziobek and Pazarbasioglu, 1997). Whether the debt relief came in the form of forgiveness, subsidized loans, or loan guarantees, the government provided equity injections to these firms.

From the viewpoint of the analysis above, bailouts occur when there is not enough private liquidity available to implement transfers of ownership quickly. Private agents anticipate that the government will supply liquidity, as indeed it does. The above examples of large-scale corporate distress—including the transition economies, Latin American economies such as Mexico, as well as Scandinavian countries—are often related to a banking crisis. This situation is analyzed next.

11.5. BANK CAPITAL, BANKING CRISES, AND BAILOUTS

We now turn to the analysis of bailouts of banking systems. In the above equilibrium, we assumed that banks were always solvent. However, there is the possibility that a troubled project could turn a solvent bank into an insolvent bank. In this section, we provide more detail about the situation of the bank. We introduce a measure of the amount of equity in a bank. As a function of how well capitalized a bank is, it may or may not behave as in the above equilibrium. In particular, a weakly capitalized bank, faced with troubled projects, may itself face the moral hazard problem of seeking to add inefficient risk. That is, there will be no incentive to liquidate projects. This means that when projects are troubled, weak banks cause a knock-on effect, where banks and entrepreneurs find it in their joint interests to engage in moral hazard.

11.5.1. Bank Capital Ratios and Bank Moral Hazard Problems

Suppose a representative bank lends to a single entrepreneur at date 0, and owes depositors an amount D at date 2. Imagine that the bank has some assets other than the projects discussed so far. These other assets have a payoff of V, where V is a random variable that will be realized at date 2. The date 2 realization will

be V_H with probability θ and V_L with probability $1 - \theta$. However, nonverifiable information about the realization of V becomes known to the banks at date 1.

We assume the risk associated with V is systematic and independent of the states of the long-term projects. In addition, we assume:

Assumption 10. *$V_H > D$, and $V_L + 1 > D$, but $V_L + L < D$. That is, when the bank receives a negative shock and the project it lent to is also in trouble, the bank is insolvent even if it receives the entire cash flow from the troubled project, L.*

Now, even though it is efficient to forgive part of the debt or to liquidate troubled projects, it can happen that the bank has a moral hazard problem itself and prefers not to liquidate troubled projects.

LEMMA 4. *If, at date 1, a bank learns that the realization of V is V_L, then the bank will not agree to renegotiate the debt contract or to liquidate a troubled project.*

Proof. See Appendix.

Previously, successful renegotiation resulted because the bank was willing to forgive some debt or to share the proceeds of liquidation with the entrepreneur. Here, by refusing to share the benefits of liquidation with the entrepreneur (via debt forgiveness), the bank removes any incentive for the entrepreneur to liquidate the project. If the entrepreneur cannot benefit by selling the project, then there is no reason to sell; adding risk is the entrepreneur's optimal strategy when the state is L. The bank engages in moral hazard by refusing to renegotiate with the borrower, thereby enticing the entrepreneur to add risk. As a result, there is a chance, δ, that the entrepreneur will be able to repay F at date 2. In that case, the bank will be able to honor its date 2 obligations to repay D to depositors.[17]

Forgiveness or liquidation is now possible only if the bank's cash flow from its other business is V_H. The bank's moral hazard problem causes an additional inefficiency in the economy. If the realization of the systematic shock is V_L, then the liquidity demand at date 1 is zero. This happens with probability $1 - \theta$. Now, the equilibrium project choices at date 0 depend on θ. We first examine the case when debt forgiveness would be feasible without the bank moral hazard problem.

17. If depositors can observe which entrepreneurs are in the low state, they can run on the banks and withdraw their deposits forcing the banks to "liquidate" (that is, the bank assets would have to be sold to the liquidity suppliers). Once they run the banks, both the bank's equity and the entrepreneur's equity are zero. Projects will be sold in the liquidation market, and the proceeds will be used to honor deposit contracts. If depositors cannot observe which entrepreneurs are in the low state, they might mistakenly run the banks with healthy projects. If depositors anticipate they might run the good banks, it may be better for them not to run any banks. These issues are discussed at length in Gorton and Huang (2001) and, for the sake of brevity, are avoided here. Here, imagine that there is deposit insurance in place (though the reasons for this are not modeled).

PROPOSITION 6. *Suppose that* $\gamma L \geq \frac{\delta}{\pi + \delta(1-\pi)}$, *i.e., debt forgiveness is feasible at date* 1. *Then:*

(*i*) *If* $\pi H + (1-\pi)[\theta L + (1-\theta)\delta TL] \geq 1 + \frac{R}{2}$, *then all entrepreneurs choose the long-term project at date* 0.

(*ii*) *If* $\pi H + (1-\pi)[\theta L + (1-\theta)\delta TL] < 1 + \frac{R}{2}$, *then all entrepreneurs choose the short-term project at date* 0.

Proof. See Appendix.

Because of the bank's moral hazard problem, renegotiation fails when the bank's state is V_L. *Ex ante,* the value of the long-term project decreases. The above proposition shows that the problem can be so severe that the long-term project can even be dominated by the short-term project. Investment in the long-term projects is then abandoned.

Next, we study how the bank moral hazard problem affects the equilibrium in the case where debt forgiveness is not feasible.

PROPOSITION 7. *Suppose that* $\gamma L < \frac{\delta}{\pi + \delta(1-\pi)}$, *i.e., debt forgiveness is not feasible at date* 1. *Then:*

(*i*) *If* $\pi H + (1-\pi)\delta TL > 1 + \frac{R}{2} + \theta(L - \delta TL)\left(1 - \frac{\delta TL - \gamma L}{R}\right)$, *then all entrepreneurs choose the long-term project at date* 0. *In this case, there will be no liquidity supply at date* 1, *and risk will be added for projects realizing the low state at date* 1.

(*ii*) *If* $\pi H + (1-\pi)[\theta L + (1-\theta)\delta TL] < 1 + \frac{R}{2}$, *then all entrepreneurs choose the short-term project at date 0.*

(*iii*) *If* $\pi H + (1-\pi)\delta TL \leq 1 + \frac{R}{2} + \theta(L - \delta TL)\left(1 - \frac{\delta TL - \gamma L}{R}\right)$, *and* $\pi H + (1-\pi)[\theta L + (1-\theta)\delta TL] \geq 1 + \frac{R}{2}$, *then there exists a measure,* α^*, *of entrepreneurs that choose the long-term project at date* 0, *and a measure,* $(1-\alpha^*)$, *of entrepreneurs that choose the short-term project at date* 0. *At date* 1 *the liquidation price is* $Q^* \in [\delta TL, L]$. α^* *and* Q^* *are defined in the Appendix.*

Proof. See Appendix.

The moral hazard problem with banks can result in troubled projects not being sold in the liquidation market at date 1. This is a kind of market failure and adds inefficiency to the economy. Recall that θ is the chance of V_H occurring. We can view θ as a measure of this additional inefficiency.

COROLLARY 2: *Social welfare is increasing in* θ.

Proof. See Appendix.

The intuition is straightforward. Higher θ means banks have more capital and are less likely to want to engage in moral hazard. Consequently, by increasing

θ, the chance that a troubled project receives equity either from the lender by forgiveness or via the liquidation market is increased, and so efficiency in the economy is improved. If banks are weaker, however, then inefficiency increases.

11.5.2. Government Bank Bailouts

Government intervention can improve efficiency when there is the potential problem of banks engaging in moral hazard. When the state is low and banks have V_L, the government has to consider not only the entrepreneurs' moral hazard problem but also the banks' moral hazard problem. Therefore the government has to offer banks a larger amount to induce banks to sell their projects to the government. The next lemma shows how much the government needs to pay the banks to bail out the troubled projects.

LEMMA 5. *Suppose the face value of the debt is* F, *and the government's tax is t at date* 2. *The government has to pay banks at least* $P' \equiv \delta(F-t)+(1-\delta)(D-V_L)$ *to remove banks' moral hazard problem.*

Proof. See Appendix.

Although it seems that the banks' moral hazard problem will make government intervention less efficient, the first-best outcome can still be reached so long as the condition $\gamma\,(\pi H+(1-\pi)L) \geq 1$ is satisfied. The intuition is as follows. Anticipating that banks will get a higher price, P', at date 1, the net interest, $(F-t)$, paid to banks can be lowered and thus the government can tax more (increase t) at date 2 to finance the subsidy at date 1. In other words, the bank moral hazard problem requires more public liquidity at date 1. However, due to the decrease in the interest paid to banks, the government has more liquidity available at date 1. So long as we have $\gamma\,(\pi H+(1-\pi)L) \geq 1$, the government has enough public liquidity to bail out all troubled projects. The first-best outcome can be attained.

If $\gamma\,(\pi H+(1-\pi)L) < 1$, government intervention cannot produce the first-best outcome because the government cannot levy enough taxes on entrepreneurs in high states. As in the case without the banks' moral hazard problem, it is possible that a liquidation market at date 1 is desirable. The following proposition characterizes the situations in which the liquidation market exists. If some projects are allowed to continue, because the government only bails out a fraction of them, then we may say that "forbearance" occurs.

PROPOSITION 8. *If* $\gamma\,(\pi H+(1-\pi)L) \geq 1$, *then government intervention can still generate the first-best outcome even if the potential banks' moral hazard problem exists. Entrepreneurs and banks sign debt contracts with a face value of*

$F = \frac{1-(1-\pi)\gamma L}{\pi}$. At date 1, the government offers to buy the loan at price $P = \frac{\delta+\pi(1-\delta)(D-V_L)}{\pi+\delta(1-\pi)}$. Banks with troubled projects sell their loans to the government and banks with high state projects retain their loans and pay a tax of $t = \frac{1-(1-\pi)\gamma L}{\pi} - \frac{1-(1-\pi)(1-\delta)(D-V_L)}{\pi+\delta(1-\pi)}$ at date 2. If $\gamma(\pi H + (1-\pi)L) < 1$, then the government can only bail out a fraction $\omega \equiv \frac{\pi((\pi+(1-\pi)\delta)\gamma H+(1-\pi)(1-\delta)(D-V_L)-1)}{(1-\pi)(\delta+\pi(1-\delta)(D-V_L)-(\pi+(1-\pi)\delta)\gamma L)}$ of the troubled long-term projects and the remaining $1-\omega$ fraction of the troubled long-term projects will either suffer from the moral hazard problem or be liquidated.

Proof. See Appendix.

The government may not be able to bail out all the low-value projects because of limited resources. Limited resources seems like a realistic assumption because of political constraints or because taxes are distortionary, so that in a larger model the social welfare-maximizing government would choose not to bail out all the low-value projects.

11.5.3. Discussion

As mentioned in the introduction, government bailouts of banking systems have recently become very common. Modem versions of the U.S. Reconstruction Finance Corporation have been used in many countries. In Mexico, for example, the Tequila crisis of 1994–1995 resulted in massive losses for Mexican banks. A public restructuring vehicle, the Trust Fund for the Protection of Bank Savings [the Fondo Bancario de Proteccion al Ahorro (FOBAPROA)], was initially used to (in part) buy loans from banks [see Honohan (no date) and Jose De Luna Martinez (2000)]. The FOBAPROA purchased nonperforming loans in an amount equal to twice the private contribution to capital, including subordinated debt, made by existing and new shareholders. The loans were purchased at book value (net of provisions) with ten-year zero coupon bonds. As Woo (2000) observes: "By purchasing the nonperforming loans from banks at book value, the FOBAPROA was essentially offering the banks free capital or a subsidy" (p. 11, footnote 14).

Argentinean banks also suffered during the Tequila crisis and the subsequent bank restructuring also involved public assistance from newly established public entities (see Augusto De la Torre, 2000). In Thailand, the Financial Institutions Development Fund—a distinct public entity—was established following the crisis of 1997. In Bulgaria, the government issued "Zunk" bond, government bonds that it used to substitute for unrecoverable bank loans. Cameroon also established a public vehicle, the Société de Recouvrement des Créances that replaced bad loans with government obligations on bank balance sheets. In Japan, there is the Financial Reconstruction Commission (see Hiroshi Nakaso, 1999). And

there are many other examples (see Andrew Sheng, 1996; William Alexander et al., 1997; Charles Enoch et al., 1999; Lindgren et al., 1999; Klingebiel, 2000).

Bailouts are not without controversy. One issue concerns whether such government safety nets generate incentive problems that we have not included in the model. For example, in our model, entrepreneurs do not have an effort choice that determines the probability of the high and low state. If they had such a choice, anticipating that the government will bail them out in the low state, entrepreneurs would shirk and free ride on other entrepreneurs' efforts (those who work hard and pay taxes to finance the bailouts). Then the *ex post* efficient government bailouts may cause an *ex ante* efficiency loss. And consequently, governments would like to commit to only bailout entrepreneurs under certain circumstances. This is an interesting and important topic, which we are pursuing.[18]

11.6. CONCLUSION

Bailouts by the government occur when the amount of the assets to be sold is so large that it would be inefficient for private agents to have hoarded liquid resources to purchase these assets in a short period of time. When the banking system is insolvent, private agents cannot readily buy the assets of the banks; it is simply not feasible since private agents lack liquidity. The government can improve welfare by creating this liquidity. However, forbearance occurs when the government cannot bail out all banks, corresponding to a situation where the government's tax capacity in the short run is too small.[19] These arguments stem from the basic idea that not all assets can be used to purchase other assets at every date.

"Liquidity" refers to the amount of readily available resources that can be used to purchase claims on projects when they are offered for sale at later dates. Not all resources can be used to buy projects. When there is a "liquidity-in-advance" constraint, the price at which claims can be sold is not just determined by the available information on their payoffs. Liquidity considerations result in prices that deviate from "efficient" market prices (i.e., the conditional expectation of the payoffs on the claim). A "liquidity discount" can arise.

18. The obvious moral hazard problem seems to be hard to detect in empirical work, suggesting that the situation is more complicated. See Gorton and Winton (2003) for a survey of the literature.

19. In our model, the constraint on the government is the amount that can be taxed at date 2. If this is too low, then not all banks can be bailed out. In reality, there may also be political constraints that prevent the government from raising taxes. For example, see Thomas Romer and Barry Weingast (1992) with regard to the U.S. thrift crisis.

At the root of the problem is the inability of private agents to buy insurance against declines in the value of their equity. Equity insurance is not available at date 0. Such insurance would have entrepreneurs with high-value projects insure entrepreneurs with low-value projects. But, this cannot occur. The incompleteness in markets raises the possibility that investment in the short-term project, what we have called hoarding, can be a desirable investment. Such investors commit to stand ready at subsequent dates to buy claims should they be offered for sale. These liquidity suppliers provide a valuable service when lender forgiveness is not optimal. But, from society's point of view it is costly to have agents engage in this activity. The government can overcome the lack of an equity insurance market by subsidizing either distressed firms or banks.

Empirically studying bailouts, and testing the model, seems like an interesting, but difficult, agenda. In reality, the issues we have discussed are complicated by the nature of the country's bankruptcy code, or lack of bankruptcy code, as well as fiscal and political considerations. Some progress is being made, however, in the form of interesting case studies. For example, in addition to the studies of bailouts and restructuring mentioned above, Enoch et al. (2002) study the transition economies, Guonan Ma and Ben S. C. Fung (2002) study China, and Mari Pangestu and Manggi Habir (2002) study Indonesia.

APPENDIX

Proof of Lemma 1:

If the bank forgives the debt at date 1, the maximum the bank can get in the low state is $f = \gamma L$. The face value of the debt, F, that was set at date 0, must be high enough such that $\pi F + (1 - \pi)\gamma L \geq 1$, or $F \geq \frac{1-(1-\pi)\gamma L}{\pi}$. But, in order for the bank to be willing to forgive debt, it must be the case that $\gamma L \geq \delta F$. Combining these two conditions, we get that $\gamma L \geq \frac{\delta}{\pi+\delta(1-\pi)}$.

Proof of Lemma 2:

Suppose the liquidation price is Q and an entrepreneur who took the short-term project has realized a cash flow of r. If $r \geq Q$, then he can afford to buy a troubled project by using his own money and no risk will be added. If $r < Q$, then he will have to borrow $Q - r$ to buy the project. He adds risk if and only if $Q - r > \gamma L$. The payoff to buying a project is $L - Q$ if risk is not added, and is $\delta TL - Q$ if risk is added.

If $Q > L$, the liquidation price is greater than the continuation value of the troubled project even if risk is not added. Therefore, no one will buy. If $Q \leq \delta TL$, buying a troubled project is profitable even if risk is added and all liquidity suppliers want to buy. If $\delta TL < Q \leq L$, then buying a troubled project is profitable only

if risk is not added. Therefore, only those liquidity suppliers with $r \geq Q - \gamma L$ will buy troubled projects.

PROOF OF LEMMA 3:
At date 1, the total measure of troubled projects is $\alpha(1-\pi)$. The total measure of liquidity at a price $Q \in [\delta TL, L]$ is $(1-\alpha)\left(1 - \frac{Q-\gamma L}{R}\right)$.

If $(1-\alpha)\left(1 - \frac{\delta TL-\gamma L}{R}\right) \leq \alpha(1-\pi)$, then liquidity demand is so high that there is no liquidation price, Q, with $Q > \delta TL$, that clears the liquidation market. Bertrand competition then drives the price down to the reservation value of δTL.

If $\alpha(1-\pi) \leq (1-\alpha)\left(1 - \frac{L-\gamma L}{R}\right)$, then there is an excess supply of liquidity at date 1. The price is at its highest level, L.

Finally, if $(1-\alpha)\left(1 - \frac{L-\gamma L}{R}\right) < \alpha(1-\pi) < (1-\alpha)\left(1 - \frac{\delta LT-\gamma L}{R}\right)$, then there exists a $Q \in (\delta LT, L)$ that clears the market. Equating liquidity demand to liquidity supply: $(1-\alpha) \times \left(1 - \frac{Q-\gamma L}{R}\right) = \alpha(1-\pi)$, results in $Q = \gamma L + R\left(1 - \frac{\alpha(1-\pi)}{1-\alpha}\right)$. All those entrepreneurs with $r \geq \delta LT - \gamma L$ get projects and will not add risk.

PROOF OF PROPOSITION 2:
Suppose the liquidation price is Q. We know that Q must be in the interval $[\delta LT, L]$. At date 0, the expected payoff to the long-term project is $\pi H + (1-\pi)Q$, which is increasing in Q; the expected payoff to the short-term project is $1 + \frac{R}{2} + (L-Q)\left(1 - \frac{Q-\gamma L}{R}\right)$, which is decreasing in Q. When liquidation price Q is δLT, the value of the long-term project reaches its minimum $\pi H + (1-\pi)\delta TL$, and the value of the short-term project reaches its maximum $1 + \frac{R}{2} + (L - \delta TL) \times \left(1 - \frac{\delta TL-\gamma L}{R}\right)$. If $\pi H + (1-\pi)\delta LT > 1 + \frac{R}{2} + (L - \delta TL)\left(1 - \frac{\delta TL-\gamma L}{R}\right)$, then taking the long-term project dominates taking the short-term project. In that case, all entrepreneurs choose the long-term project at date 0. Suppose $\pi H + (1-\pi)\delta TL \leq 1 + \frac{R}{2} + (L - \delta TL) \times \left(1 - \frac{\delta TL-\gamma L}{R}\right)$. Recall that we have assumed that $\pi H + (1-\pi)L > 1 + R/2$. Therefore, there exists a unique $Q^* \in [\delta TL, L]$ such that $\pi H + (1-\pi)Q^* = 1 + \frac{R}{2} + (L - Q^*) \times \left(1 - \frac{Q^*-\gamma L}{R}\right)$. Solving for Q^*, we get:

$$Q^* = \frac{(2-\pi)R + (1-\gamma)L - \sqrt{((2-\pi)R + (1-\gamma)L)^2 - 4\left(R\left(1 + \frac{R}{2} - \pi H + L\right) + \gamma L^2\right)}}{2}.$$

To obtain α^*, the secondary liquidation market must be cleared at date 1, i.e., $(1-\alpha) \times \left(1 - \frac{Q^*-\gamma L}{R}\right) = \alpha(1-\pi)$. Solving this equation gives: $\alpha^* = \frac{R + \gamma L - Q^*}{(2-\pi)R + \gamma L - Q^*}$.

PROOF OF PROPOSITION 3 AND COROLLARY 1:

(1) Define

$$G(\pi,\delta,R,Q^*)=\pi H+(1-\pi)Q^*-1-\frac{R}{2}-(L-Q^*)\left(1-\frac{Q^*-\gamma L}{R}\right).$$

$$\frac{\partial Q^*}{\partial R}=-\frac{\partial G/\partial R}{\partial G/\partial Q^*}=\frac{\frac{1}{2}+\frac{(L-Q^*)(Q^*-\gamma L)}{R^2}}{(1-\pi)+\left(1-\frac{Q^*-\gamma L}{R}\right)+\frac{L-Q^*}{R}}>0;$$

$$\frac{\partial Q^*}{\partial R}=-\frac{\partial G/\partial \pi}{\partial G/\partial Q^*}=-\frac{H-Q^*}{(1-\pi)+\left(1-\frac{Q^*-\gamma L}{R}\right)+\frac{L-Q^*}{R}}<0;$$

$$\frac{\partial Q^*}{\partial \delta}=-\frac{\partial G/\partial \delta}{\partial G/\partial Q^*}=-\frac{\frac{(L-Q^*)L(T-1)}{R(1-\delta^2)}}{(1-\pi)+\left(1-\frac{Q^*-\gamma L}{R}\right)+\frac{L-Q^*}{R}}<0.$$

(2) To prove that social welfare is increasing in R and decreasing in δ, we use the expected payoff to entrepreneurs who take the long-term project. Define $W=\pi H+(1-\pi)Q^*$. Then: $\frac{\partial W}{\partial R}=(1-\pi)\frac{\partial Q^*}{\partial R}>0$, and $\frac{\partial W}{\partial \delta}=(1-\pi)\frac{\partial Q^*}{\partial \delta}<0$. To prove that social welfare is increasing in π, we use the expected payoff to entrepreneurs who take the short-term project. Define $W=1+\frac{R}{2}+(L-Q^*)\left(1-\frac{Q^*-\gamma L}{R}\right)$. Then, $\frac{\partial W}{\partial \pi}=-\left(\frac{(L-Q^*)}{R}+1-\frac{Q^*-\gamma L}{R}\right)\frac{\partial Q^*}{\partial \pi}>0$.

PROOF OF PROPOSITION 4:

Suppose the face value of the debt is F and the government levies t on each bank at date 2. $F-t$ is the net payment a bank receives from a loan in the high state. When the government buys the loans from banks at date 1, the minimum price it has to pay to the banks is $\delta(F-t)$, which is the payoff banks with low-state projects can get if risk is added. And, in order to remove the incentive to engage in moral hazard, the government has to forgive the face value of the debt to less than or equal to γL. Therefore, $\delta(F-t)-\gamma L$ is the necessary subsidy to bail out a troubled project. The government has a budget constraint condition: $\pi t\geq(1-\pi)(\delta(F-t)-\gamma L)$. Banks have to break even, so we have $\pi(F-t)+(1-\pi)\delta(F-t)\geq 1$. Finally, the face value of the debt cannot be too high; otherwise even entrepreneurs in the high state will add risk. So, we must have $F\leq\gamma H$. Combining these conditions, we obtain $\gamma(\pi H+(1-\pi)L)\geq 1$. The equilibrium is the solution to the following system of equations: $P=\delta(F-t)$, $\pi t=(1-\pi)(P-\gamma L)$, $\pi(F-t)+(1-\pi)P=1$.

Solving for the equilibrium gives us: $F = \frac{1-(1-\pi)\gamma L}{\pi}$, $P = \frac{\delta}{\pi+\delta(1-\pi)}$, and $t = \frac{1-(1-\pi)\gamma L}{\pi} - \frac{1}{\pi+\delta(1-\pi)}$.

PROOF OF PROPOSITION 5:

When $\gamma(\pi H + (1-\pi)L) < 1$, the first-best outcome characterized in Proposition 4 is not feasible because the government lacks enough resources to bail out all distressed long-term projects. The government chooses the loan price, P, tax on high projects, t, and bails out a fraction, ω, of the low state projects. We first suppose that there is private liquidity supply [i.e., Q solves $\pi H + (1-\pi)[\omega L + (1-\omega)Q] = 1 + \frac{R}{2} + (L-Q)\left(1 - \frac{Q-\gamma L}{R}\right)$] and show that the government's optimal choice is to maximize ω.

Define $J(\pi, \delta, R, Q) = \pi H + (1-\pi)[\omega L + (1-\omega)Q] = 1 + \frac{R}{2} + (L-Q)\left(1 - \frac{Q-\gamma L}{R}\right)$.

$$\frac{\partial Q}{\partial \omega} = -\frac{\partial J/\partial \omega}{\partial J/\partial Q} = -\frac{(1-\pi)(L-Q)}{(1-\pi)(1-\omega) + \left(1 - \frac{Q-\gamma L}{R}\right) + \frac{L-Q}{R}} < 0.$$

In equilibrium the expected payoff to entrepreneurs who take the long-term project and the expected payoff to entrepreneurs who take the short-term project are the same, which is equal to $\pi H + (1-\pi)[\omega L + (1-\omega)Q]$. We show it is increasing in ω.

Define $W = \pi H + (1-\pi)[\omega L + (1-\omega)Q]$.

$$\frac{\partial W}{\partial R} = (1-\pi)\left[(L-Q) + (1-\omega)\frac{\partial Q}{\partial \omega}\right]$$

$$= (1-\pi)(L-Q)\frac{\left(1 - \frac{Q-\gamma L}{R}\right) + \frac{L-Q}{R}}{(1-\pi)(1-\omega) + \left(1 - \frac{Q-\gamma L}{R}\right) + \frac{L-Q}{R}} > 0.$$

By constraint (4), in order to maximize ω, the government needs to minimize P. To satisfy constraint (2), P will be set to equal $\delta(F-t)$, which is the expected payoff banks receive if risk is added to low state projects, and is also the expected payoff banks receive in case the projects are sold because the bargaining power is in the hands of entrepreneurs. Substituting $P = \delta(F-t)$ into constraint (3), we get $\pi(F-t) + (1-\pi)\delta(F-t) = 1$, or $t = F - \frac{1}{\pi+(1-\pi)\delta}$.

Since the government has to subsidize each troubled project by the amount $\delta(F-t) - \gamma L$, the fraction of troubled projects that can be subsidized is equal to $\frac{(\pi+(1-\pi)\delta)F-1}{(1-\pi)(\delta F-\gamma L)}$. This fraction is maximized when F is equal to γH. The maximum is equal to $\omega^* \equiv \frac{\pi((\pi+(1-\pi)\delta)\gamma H-1)}{(1-\pi)(\delta-(\pi+(1-\pi)\delta)\gamma L)}$. Note that ω is always less than one because we have $\gamma(\pi H + (1-\pi)L) < 1$. The remaining $(1-\omega)$ projects

cannot be bailed out by the government. Risk will be added to these projects if they are not liquidated. Then the question is whether there are liquidity suppliers in the secondary market. We need to check whether a deviation to the short-term project is profitable if all other entrepreneurs take the long-term project.

Suppose $\pi H + (1-\pi)[\omega^* L + (1-\omega^*)\delta TL] > 1 + \frac{R}{2} + (L - \delta TL)$, i.e., there is no entrepreneur willing to supply liquidity. Can the government lower ω to improve welfare? The answer is no. Suppose the government lowers ω to a certain level such that some entrepreneurs are willing to invest in the short-term project. According to the proof above, the governments will set ω as high as possible, so the maximum point is reached when $Q = \delta TL$. But when $Q = \delta TL$, welfare must be less than $\pi H + (1-\pi)[\omega^* L + (1-\omega^*)\delta TL]$ because ω cannot be larger than ω^*.

Similar to the proof of Proposition 2, we can solve for Q^* and α^*. Since entrepreneurs must be indifferent between taking the short-term project and taking the long-term project at date 1, the liquidation price Q^* must satisfy $\pi H + (1-\pi)[\omega L + (1-\omega)Q] = 1 + \frac{R}{2} + (L - Q)\left(1 - \frac{Q - \gamma L}{R}\right)$. Solving for Q^*, we get:

$$Q^* = \frac{(1+(1+\omega)(1-\pi))R + (1+\gamma)L - \sqrt{\begin{array}{l}((1+(1-\omega)(1-\pi))R + (1+\gamma)L)^2 \\ \quad -4\left(R\left(1+\frac{R}{2} - \pi H + L(1-\omega(1-\pi))\right) + \gamma L^2\right)\end{array}}}{2}.$$

Finally, substituting Q^* into the market-clearing condition at date 1 gives: $\alpha(1-\pi)(1+\omega) = (1-\alpha)\left(1 - \frac{Q^* - \gamma L}{R}\right)$. Solving this equation gives us $\alpha^* = \frac{R + \gamma L - Q^*}{(1+(1-\pi)(1+\omega))R + \gamma L - Q^*}$.

PROOF OF LEMMA 4:

The bank can get at most L from forgiveness or liquidation. In these cases, its equity is negative. Therefore, the bank will not forgive debt or sell the project, forcing continuation with risk added. In that case, there is still some hope that the bank will be solvent at date 2.

PROOF OF PROPOSITION 6:

If the systematic shock turns out to be high, i.e., V is equal to V_H, then there will be debt forgiveness and the value of the project is L in the low state. If the systematic shock turns out to be low, i.e., V is equal to V_L, then there will be no debt forgiveness by Lemma 4 and the value of the project is δTL in the low state. Therefore the expected payoff from taking the long-term project is $\pi H + (1-\pi)[(1-\theta)\delta TL + \theta L]$. On the other hand, if an entrepreneur takes a short-term project, there is no chance for him to buy a troubled project and hence the expected payoff is $1 + \frac{R}{2}$. At date 0

entrepreneurs make investment decisions by comparing these two expected values.

PROOF OF PROPOSITION 7:
If $\pi H+(1-\pi)\delta TL > 1+\frac{R}{2}+\theta(L-\delta TL)\left(1-\frac{\delta TL-\gamma L}{R}\right)$, then the return to the long-term project dominates that of the short-term project even if the date 1 liquidation price is at its reservation value δTL. Therefore, no entrepreneur invests in the short-term project at date 0.

If $\pi H+(1-\pi)[(1-\theta)\delta TL+\theta L] \leq 1+\frac{R}{2}$, then even if the troubled project can be sold for its maximum value of L in the liquidation market, θ is so low that, *ex ante,* the payoff from a long-term project is less than the payoff from a short-term project. No entrepreneur chooses the long-term project at date 0.

Suppose $\pi H+(1-\pi)\delta TL \leq 1+\frac{R}{2}+\theta(L-\delta TL)\left(1-\frac{\delta TL-\gamma L}{R}\right)$ *and* $\pi H+(1-\pi)\times[(1-\theta)\delta TL+\theta L] > 1+\frac{R}{2}$. Then there is a unique liquidation price, Q^*, that solves the equation: $\pi H+(1-\pi)[(1-\theta)\delta TL+\theta Q] = 1+\frac{R}{2}+\theta(L-Q)\left(1-\frac{Q-\gamma L}{R}\right)$. Solving this equation, we get

$$Q^*=\frac{(2-\pi)R+(1+\gamma)L-\sqrt{((2-\pi)R+(1+\gamma)L)^2-4\left(\frac{R}{\theta}\left(1+\frac{R}{2}-\pi H-(1-\pi)(1-\theta)\delta TL+\theta L\right)+\gamma L^2\right)}}{2}.$$

At Q^* an individual entrepreneur is indifferent between choosing the long-term project or the short-term project at date 0. Equilibrium project choice must be consistent with Q^*, so α^* is the solution to the equation: $(1-\alpha)\left(1-\frac{Q^*-\gamma L}{R}\right)=\alpha(1-\pi)$. Solving for α^*, we get $\alpha^*=\frac{R+\gamma L-Q^*}{(2-\pi)R+\gamma L-Q^*}$.

PROOF OF COROLLARY 2:
When debt forgiveness is feasible, social welfare is equal to

$$\max\left\{\pi H+(1-\pi)[\theta L+(1-\theta)\delta TL], 1+\frac{R}{2}\right\},$$

which is increasing in θ. Suppose debt forgiveness is not feasible. If all entrepreneurs take the short-term project or all entrepreneurs take the long-term project, there is no liquidation market. Social welfare is either $\pi H+(1-\pi)\delta TL$ or $1+\frac{R}{2}$. If entrepreneurs choose different projects and there is a liquidation market at date 1, social welfare is equal to $\pi H+(1-\pi)[\theta Q+(1-\theta)\delta TL]$. Let $W(\theta)\equiv\pi H+(1-\pi)[\theta Q+(1-\theta)\delta TL]$. We have $W'(\theta)=(1-\pi)Q-\delta TL+\theta\frac{\partial Q}{\partial\theta}>0$.

PROOF OF LEMMA 5:
To make V_L banks sell the troubled projects, the government has to pay a price, P', such that $V_L + P' - D \geq \delta(V_L + F - t - D)$. Rearranging the terms, we get $P' \geq \delta(F - t) + (1 - \delta)(D - V_L)$.

PROOF OF PROPOSITION 8:
Suppose the face value of the debt is F and the government levies tax t on each bank at date 2. By Lemma 5, when the government buys the claims from banks at date 1, the minimum price it has to pay to the banks is $\delta(F - t) + (1 - \delta)(D - V_L)$. And in order to remove the entrepreneurs' incentives to engage in moral hazard, the government has to forgive the face value of the debt to less than or equal to γL. The government has a budget constraint: $\pi t \geq (1 - \pi)(\delta(F - t) + (1 - \delta)(D - V_L) - \gamma L)$. Banks have to break even, so we have $\pi(F - t) + (1 - \pi)(\delta(F - t) + (1 - \delta)(D - V_L)) \geq 1$. Finally, the face value of the debt cannot be too high because otherwise even entrepreneurs in the high state will add risk. So, we must have $F \leq \gamma H$. Combining these conditions, we obtain $\gamma(\pi H + (1 - \pi)L) \geq 1$. The equilibrium is the solution to the following system of equations: $P = \delta(F - t) + (1 - \delta)(D - V_L), \pi t = (1 - \pi)(P - \gamma L), \pi(F - t) + (1 - \pi)P = 1$. Solving for the equilibrium gives us: $F = \frac{1-(1-\pi)\gamma L}{\pi}, P = \frac{\delta + \pi(1-\delta)(D-V_L)}{\pi + \delta(1-\pi)}$, and $t = \frac{1-(1-\pi)\gamma L}{\pi} - \frac{1-(1-\pi)(1-\delta)(D-V_L)}{\pi + \delta(1-\pi)}$.

Now suppose $\gamma(\pi H + (1 - \pi)L) < 1$. Similar to the Proof of Proposition 5, the government chooses the largest possible ω to maximize its objective function. This is achieved by setting: $\pi(F - t) + (1 - \pi)(\delta(F - t) + (1 - \delta)(D - V_L)) = 1, P = \delta(F - t) + (1 - \delta)(D - V_L)$, and $\pi t = (1 - \pi)\omega(P - \gamma L)$, which correspond to the banks' break-even condition, incentive compatibility condition, and the government's budget constraint, respectively. Moreover, F takes its highest possible value, γH. The maximum fraction is equal to $\omega \equiv \frac{\pi((\pi + (1-\pi)\delta)\gamma H + (1-\pi)(1-\delta)(D-V_L) - 1)}{(1-\pi)(\delta + \pi(1-\delta)(D-V_L) - (\pi + (1-\pi)\delta)\gamma L)}$. Note that ω is always less than one because we have $\gamma(\pi H + (1 - \pi)L) < 1$. The remaining $(1 - \omega)$ projects cannot be bailed out by the government and risk will be added to these projects if they are not liquidated.

The next question is whether there are liquidity suppliers in the secondary market. We need to check whether deviation to the short-term project is profitable if all other entrepreneurs take the long-term project. If all other entrepreneurs take the long-term project, and one entrepreneur takes the short-term project, then this entrepreneur is the sole liquidity supplier at date 1 and he can buy a troubled project at price δTL. His expected payoff from taking the short-term project is: $1 + \frac{R}{2} + \theta(L - \delta TL)\left(1 - \frac{\delta TL - \gamma L}{R}\right)$. If this entrepreneur does not deviate, his expected payoff is: $\pi H + (1 - \pi)[\omega L + (1 - \omega)\delta TL]$. Therefore, there are liquidity suppliers if and only if $\pi H + (1 - \pi)[\omega L + (1 - \omega)\delta TL] \leq 1 + \frac{R}{2} + \theta(L - \delta TL)\left(1 - \frac{\delta TL - \gamma L}{R}\right)$.

The remainder of the proof is similar to the Proof of Proposition 2. Since entrepreneurs must be indifferent between taking the short-term project and taking the long-term project at date 1, the liquidation price Q^* must satisfy:

$$\pi H + (1-\pi)[\omega L + (1-\omega)(\theta Q + (1-\theta)\delta TL)] = 1 + \frac{R}{2} + \theta(L-Q)\left(1 - \frac{Q - \gamma L}{R}\right).$$

Solving for Q^*, we get:

$$Q^* = \frac{(1+(1+\omega)(1+\pi))R + (1+\gamma)L - \sqrt{\begin{array}{l}((1+(1-\omega)(1-\pi))R + (1+\gamma)L)^2 \\ \quad -4\left(\frac{R}{\theta}\left(1+\frac{R}{2} - \pi H + L(\theta - \omega(1-\pi))\right.\right. \\ \quad \left.\left. -(1-\pi)(1-\omega)(1-\theta)\delta TL\right) + \gamma L^2\right).\end{array}}}{2}$$

Once we get Q^*, we plug Q^* into the market-clearing condition at date 1: $\alpha(1-\pi)(1-\omega) = (1-\alpha)\left(1 - \frac{Q-\gamma L}{R}\right)$. Solving this equation gives us: $\alpha^* = \frac{R+\gamma L - Q^*}{(1+(1-\pi)(1-\omega))R + \gamma L - Q^*}$.

REFERENCES

Alexander, William; Davis, Jeffrey; Ebrill, Liam and Lindgren, Carl-John. *Systemic bank restructuring and macroeconomic policy.* Washington, DC: International Monetary Fund, 1997.

Allen, Franklin and Gale, Douglas. "Optimal Financial Crises." *Journal of Finance,* August 1998, *53*(4), pp. 1245–84.

Barth, James. *The great savings and loan debacle.* Washington, DC: AEI Press, 1991.

Barth, James and Bartholomew, Philip. "The Thrift Industry Crisis: Revealed Weaknesses in the Federal Deposit Insurance System," in James Barth and R. Dan Brumbaugh, Jr., eds., *The reform of federal deposit insurance: Disciplining the government and protecting taxpayers.* New York: HarperBusiness, 1992, pp. 36–116.

Bean, Mary L.; Hodge, Martha; Ostermiller, William; Spaid, Mike and Stockton, Steve. "Executive Summary: Resolution and Asset Disposition Practices in Federal Deposit Insurance Corporation," in *Managing the crisis: The FDIC and RTC experience, 1980–1994.* Washington, DC: FDIC, 1998, pp. 3–52.

Bhattacharya, Sudipto and Gale, Douglas M. "Preference Shocks, Liquidity, and Central Bank Policy," in William A. Barnett and Kenneth J. Singleton, eds., *New approaches to monetary economics: Proceedings of the second international symposium in economic theory and econometrics.* Cambridge, MA: Cambridge University Press, 1987, pp. 69–88.

Brumbaugh, R. Dan, Jr. *Thrifts under siege.* Cambridge, MA: Ballinger Publishing Company, 1988.

Calomiris, Charles. "Is the Discount Window Necessary? A Penn Central Perspective?" *Federal Reserve Bank of St. Louis Review*, May/June 1994, pp. 31–55.

Caprio, Gerard and Klingebiel, Daniela. "Bank Insolvencies: Cross Country Experience." World Bank Policy Research Working Paper No. 1620, 1996.

———. "Episodes of Systemic and Borderline Financial Crises." Mimeo, World Bank, 1999.

Claessens, Stijn; Djankov, Simeon and Klingebiel, Daniela. "Financial Restructuring in East Asia: Halfway There?" World Bank Financial Sector Discussion Paper No. 3, 1999.

Daniel, James A. "Fiscal Aspects of Bank Restructuring." Working paper, International Monetary Fund, 1997.

De la Torre, Augusto. "Resolving Bank Failures in Argentina: Recent Developments and Issues." World Bank Policy Research Working Paper No. 2295, 2000.

De Luna Martinez, Jose. "Management and Resolution of Banking Crises: Lessons from the Republic of Korea and Mexico." World Bank Discussion Paper No. 413, 2000.

Diamond, Douglas W. "Liquidity, Banks, and Markets." *Journal of Political Economy*, October 1997, *105*(5), pp. 928–56.

Diamond, Douglas W. and Dybvig, Philip H. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy*, June 1983, *91*(3), pp. 401–19.

Diamond, Douglas W. and Rajan, Raghuram. "Liquidity Risk, Liquidity Creation, and Financial Fragility: A Theory of Banking." *Journal of Political Economy*, April 2001, *109*(2), pp. 287–327.

———."Bank Bailouts and Aggregate Liquidity." Unpublished manuscript, University of Chicago, 2002.

Dziobek, Claudia and Pazarbasioglu, Ceyla. "Lessons from Systemic Bank Restructuring: A Survey of 24 Countries." International Monetary Fund Working Paper No. 97/161, December 1997.

———."Lessons From Systemic Bank Restructuring." International Monetary Fund *Economic Issues,* No. 14, 1998.

Eisfeldt, Andrea. "Endogenous Liquidity in Asset Markets." Unpublished manuscript, Kellogg School, Northwestern University, 2002.

Enoch, Charles; Garcia, Gillian and Sundararajan, V. "Recapitalizing Banks with Public Funds: Selected Issues." International Monetary Fund Working Paper No. 99/139, 1999.

Enoch, Charles; Gulde, Anne-Marie and Hardy, Daniel. "Banking Crises and Bank Resolution: Experiences in Some Transition Economies." International Monetary Fund Working Paper No. 02/56, 2002.

Goodhart, Charles A. E. *The central bank and the financial system.* London: Macmillan, 1995.

Gorton, Gary B. and Huang, Lixin. "Bank Panics and the Endogeneity of Central Banking." Working paper, Wharton School, 2001.

Gorton, Gary B. and Kahn, James. "The Design of Bank Loan Contracts." *Review of Financial Studies*, Summer 2000, *13*(2), pp. 331–54.

Gorton, Gary B. and Pennacchi, George. "Financial Intermediaries and Liquidity Creation." *Journal of Finance,* March 1990, *45*(1), pp. 49–71.

Gorton, Gary B. and Winton, Andrew. "Financial Intermediation," in George Constantinides, Milton Harris, and René Stulz, eds., *Handbook of the economics of finance.* Amsterdam: North-Holland, 2003, pp. 431–552.

Grossman, Sanford J. "An Analysis of the Implications for Stock and Futures Price Volatility of Program Trading and Dynamic Hedging Strategies." *Journal of Business,* July 1988, *61*(3), pp. 275–98.

Hawkins, John and Turner, Philip. "Bank Restructuring in Practice: An Overview." Bank for International Settlements Policy Paper No. 6, 1999.

Holmström, Bengt and Tirole, Jean. "Private and Public Supply of Liquidity." *Journal of Political Economy,* February 1998, *106*(1), pp. 1–40.

Honohan, Patrick. "Recapitalizing Banking Systems: Implications for Incentives and Fiscal and Monetary Policy." World Bank Working Paper No. 2540 (no date).

Honohan, Patrick and Klingebiel, Daniela. "Controlling Fiscal Costs of Bank Crises." World Bank Working Paper No. 2441, 2000.

Hopenhayn, Hugo A. and Werner, Ingrid M. "Information, Liquidity, and Asset Trading in a Random Matching Game." *Journal of Economic Theory,* February 1996, *68*(2), pp. 349–79.

Iacocca, Lee and Novak, William, *Iacocca: An autobiography.* New York: Bantam Books, 1986.

Kane, Edward. "Dangers of Capital Forbearance: The Case of the FSLIC and 'Zombie' S&Ls." *Contemporary Policy Issues,* January 1984, *5*(1), pp. 77–83.

——. *The S&L mess: How did it happen?* Washington, DC: Urban Institute Press, 1989.

Kane, Edward J. and Yu, Min-Teh. "Opportunity Cost of Capital Forbearance during the Final Years of the FSLIC Mess." *Quarterly Review of Economics and Finance,* Fall 1996, *36*(3), pp. 271–90.

Klingebiel, Daniela. "The Use of Asset Management Companies in the Resolution of Banking Crises: Cross-Country Experience." World Bank Policy Research Working Paper No. 2284, 2000.

Kyle, Albert S. "Continuous Auctions and Insider Trading." *Econometrica,* November 1985, *53*(6), pp. 1335–55.

Lindgren, Carl-Johan; Baliño, Tomás J. T.; Enoch, Charles; Gulde, Anne-Marie; Quintyn, Marc and Teo, Leslie. "Financial Sector Crisis and Restructuring: Lessons From Asia." International Monetary Fund Occasional Paper No. 188, 1999.

Lowenstein, Roger. *When genius failed.* New York: Random House, 2000.

Lummer, Scott and McConnell, John. "Further Evidence on the Bank Lending Process and the Capital-Market-Responses to Bank Loan Agreements." *Journal of Financial Economics,* November 1989, *25*(1), pp. 99–122.

Ma, Guonan and Fung, Ben S. C. "China's Asset Management Corporations." Bank for International Settlements Working Paper No. 115, 2002.

Mason, Joseph. "Reconstruction Finance Corporation Assistance to Financial Institutions and Commercial & Industrial Enterprise in the U.S. Great Depression, 1932–1937," in Stijn Claessens, Simeon Djankov, and Ashoka Mody, eds., *Resolution of financial distress.* Washington, DC: World Bank Press, 2001, pp. 167–204.

Nakaso, Hiroshi. "Recent Banking Sector Reforms in Japan." *Economic Policy Review,* Federal Reserve Bank of New York, July 1999, 5(2), pp. 1–7.

Olson, James Stuart. *Herbert Hoover and the Reconstruction Finance Corporation, 1931–1933.* Ames, IA: Iowa State University Press, 1977.

Pangestu, Mari and Habir, Manggi. "The Boom, Bust, and Restructuring of Indonesian Banks." International Monetary Fund Working Paper No. 02/66, 2002.

Romer, Thomas and Weingast, Barry. "Political Foundations of the Thrift Debacle," in James Barth and R. Dan Brumbaugh, Jr., eds., *The reform of federal deposit insurance: Disciplining the government and protecting tax payers.* New York: HarperBusiness, 1992, pp. 167–202.

Sheng, Andrew. *Bank restructuring: Lessons from the 1980s.* Washington, DC: World Bank, 1996.

Shleifer, Andrei and Vishny, Robert. "Liquidation Values and Debt Capacity: A Market Equilibrium Approach." *Journal of Finance,* September 1992, *47*(4), pp. 1343–66.

Stone, Mark. "Corporate Sector Restructuring: The Role of Government in Times of Crisis." *Economic Issues,* International Monetary Fund, June 2002, *31.*

Todd, Walker. "History of and Rationales for the Reconstruction Finance Corporation." Federal Reserve Bank of Cleveland. *Economic Review,* Quarter 4, 1992, pp. 22–35.

White, Lawrence. *The S&L debacle: Public policy lessons for bank and thrift regulation.* New York: Oxford University Press, 1991.

Woo, David. "Two Approaches to Resolving Nonperforming Assets During Financial Crises." International Monetary Fund Working Paper No. 00/03, 2000.

PART III

WHAT DO BANKS DO?

12

The Design of Bank Loan Contracts

GARY B. GORTON AND JAMES KAHN*

Empirical work strongly suggests that bank loans are different from corporate bonds.[1] This evidence has spawned a number of hypotheses about exactly what banks do to make themselves valuable. These theories have stressed various kinds of screening and monitoring of borrowers. In this chapter we argue that the interesting and valuable functions of banks occur between the time they make a loan and collect repayment. We focus on banks' ability to renegotiate credit terms with borrowers, and on the tight link between that renegotiation and monitoring. Our model shows how the unique characteristics of bank loans emerge endogenously to enhance efficiency. These characteristics include seniority (i.e., the bank has first claim on the assets of the borrower in the event of default); an option for the bank to liquidate the loan at any time (perhaps in the form of very

* This is a revised version of a previous article with a slightly different title. Thanks to Mark Carey, Mathias Dewatripont, Douglas Diamond, Oliver Hart, Paul Milgrom, Raghuram Rajan, and David Webb for discussions and to Nils Gottfries, Michel Habib, Leonard Nakamura, an anonymous referee, and seminar participants at the University of Chicago, University of Illinois, Board of Governors, Johns Hopkins, Wayne State, ECARE, the CEPR Meeting at Toulouse, the University of Stockholm, the Penn Macro Lunch Group and the Penn Finance Lunch Group, for suggestions. The views expressed are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of New York or the Federal Reserve System.

1. For example, James (1987) finds a positive and significant abnormal stock response to firms announcing the signing of bank loan agreements. Also see Lummer and McConnell (1989). Hoshi, Kashyap, and Scharfstein (1990) find that Japanese firms in financial distress that are members of a "main-bank" coalitions (keiretsu) invest and sell more after the onset of distress than do distressed firms that are not members of a bank coalition. Other evidence includes Gilson, John, and Lang (1990) and Slovin, Sushka, and Polonchek (1993).

tight covenants); and an initial loan rate set not to price the risk of default, but to minimize subsequent costs associated with moral hazard and renegotiation. As a consequence of this last feature, initial loan pricing may involve transfers—either from the borrower to the bank (in the form of up-front fees, compensating balance requirements, or the purchase of other bank services), or from the bank to the borrower with the bank providing underpriced services to the borrower. The model also predicts that firm risk is endogenous and state dependent, and that renegotiated interest rates on the debt need not be monotone in firm risk.

While loans and bonds are both debt contracts, we take the defining feature of bank loans to be that they are typically held by a single creditor. We will argue that this feature makes "monitoring" of the borrower both feasible and desirable. Bonds, on the other hand, are (by definition) held by dispersed creditors for whom monitoring and renegotiation are relatively costly because of free-rider problems.[2] We incorporate this distinction in a model of renegotiation between a borrower and a bank lender. Renegotiation of the contract terms is triggered by the arrival of new information that may lead the borrower to add inefficient risk to the project (i.e., asset substitution) absent changes in the terms of the loan. There is also the potential for moral hazard on the part of the bank since the bank may "hold up" the borrower by (credibly) threatening to liquidate the borrower's project, thereby extracting a higher interest rate.

The interplay between the two moral hazard problems leads to a number of outcomes to renegotiation. The bank may liquidate the project, raise the interest rate, forgive some of the debt, or stay with the status quo. "Monitoring" the borrower can be interpreted to mean liquidating inefficient projects and renegotiating lower interest rates to prevent borrower risk taking. But we show that in renegotiation the bank is not always successful in preventing the borrower from taking on additional risk. Sometimes the bank allows a borrower to continue with the project even though the borrower chooses to add risk to the project. In equilibrium, the variance of the value of the borrowing firm is therefore endogenously time and state dependent. Because the bank can only succeed in preempting risky behavior in the moderately distressed cases by writing off some of the debt or lowering the rate, renegotiation results in renegotiated interest rates that are not monotone in borrower quality: the healthiest borrowers are left alone, the moderately distressed are granted concessions, while the most distressed are forced to submit to harsher terms.

The contract design problem involves a number of considerations, each of which we address. First, there is the question of whether renegotiation is

2. Our model is consistent with any secured debt-holder who has sufficient bargaining power to renegotiate with a borrower (and we do not take a stand on how large a position this requires). Typically banks are single lenders, making renegotiation practical. Kahan and Tuckman (1993) argue that firms do have mechanisms at their disposal to negotiate with decentralized bondholders, but they are potentially costly to shareholders.

desirable. In other words, is it efficient for the borrower to obtain funds from a bank, as opposed to obtaining funds from agents who cannot renegotiate? Answering this question involves comparing the outcomes of obtaining funds from a single lender, such as a bank, to the alternative of issuing bonds to dispersed lenders. Issuing bonds commits the firm and its creditors not to renegotiate. The second design issue concerns the contract with the bank, if funds are obtained from a bank. Here the question is whether the contract should include a provision which allows the bank to ask for the collateral prior to maturity of the loan (even if the borrower has not missed a payment). We assume that the contract can feasibly include the liquidation option which allows the bank to "call the loan" at any time, and we ask whether it is optimal to include this provision.

If the liquidation option is included, then the third contract design consideration involves the specification of the initial contract form, considering that both parties know that at an interim date the contract can be renegotiated upon the arrival of new information. While we assume that if the project continues at the interim date it must do so under a debt contract that matures at a final date, this does not determine the optimal form of the initial contract, since the borrower and the lender know that any initial contract will subsequently be renegotiated. The outcome of the renegotiation has efficiency considerations, since some projects will be liquidated by the bank, while others will become riskier (when borrowers add risk). The social gain from bank loans comes from the enhanced ability to thwart inefficient risk taking and to liquidate bad projects. Because the bank may liquidate too frequently, however, the net value of bank loans rests on the costs of excessive liquidation being small relative to the costs of excessive continuation. We show how the terms of the initial contract affect the renegotiation outcome by allocating bargaining power between borrower and lender to minimize inefficient risk taking.

Our model identifies a unique role for bank loans that is independent of pricing default risk. The initial equilibrium interest rate on loans does not primarily reflect a default premium. Rather it is the rate that results ex ante in minimal expected asset substitution by borrowers following renegotiation. The loan is certain to be renegotiated, and the outcome of bargaining between the two parties is partly determined by the bank's threat to liquidate. But the credibility of this threat depends in part on the amount owed to the bank. Intuitively, the amount owed must be high enough so that the bank will not be overly tempted to hold up the borrower for higher payments and thereby induce excessive risk taking, but not so high that the bank would be insufficiently willing to forgive some of the debt in order to discourage excessive risk taking. Given such considerations, there is no guarantee that the loan rate that minimizes these expected agency costs will result in zero expected profits for lenders. Consequently, competition by banks can result in nonlinear pricing arrangements for loans such as origination fees or cross-subsidization with other products, as are

often observed. Previous explanations of the structure of bank loan pricing have relied on screening in asymmetric information environments [e.g., Thakor and Udell (1987)].

Our results are related to the literature on the role of banks, including Sharpe (1990), Rajan (1992), and Detragiache (1994). In the models of Sharpe and Rajan, banks learn private information about borrowers and are able to exploit this information to hold up borrowers. We include this moral hazard on the part of the banks and, in addition, include moral hazard by the borrower. In Detragiache's model renegotiation is beneficial, but can lead to ex ante risk taking by the borrower. Her focus is on alternative bankruptcy regimes.

Our article is also related to the literature on the role of banks as ex post monitors, which views banks' primary role as verifying reported (and otherwise unobservable) output in settings with costly state verification [e.g., Diamond (1984)]. This theory cannot explain observed interaction between banks and borrowers during the life of the contract. Moreover, the role of banks as ex post monitors suggests that banks should be junior claimants (and perhaps equity claimants) because their incentive to monitor would then be strongest.[3] Fama (1985) argues that this is the case. But in fact, banks are typically senior, secured claimants. It seems difficult to reconcile this feature of bank loans with the bank's role as ex post monitor. Our model addresses this issue.

The model is specified in Section 12.1 Section 12.2 provides preliminary results and definitions of payoffs. Section 12.3 looks at the renegotiation and liquidation decisions predicted by the model. Section 12.4 examines the initial pricing of the loan and the role of debt. Section 12.5 discusses the results, and Section 12.6 contains some final remarks.

12.1. THE BORROWING AND LENDING ENVIRONMENT

There are four dates, $t = 0, 1, 2, 3$, in the model economy and two representative risk-neutral agents: a borrowing firm and a lender (which we will call the "bank"). A summary of the model is as follows. The borrowing firm has a project which requires some external financing: at date $t = 0$ the firm obtains funding from a competitive bank. The funding is governed by a contract that matures at date $t = 2$. At $t = 1$, before the contract matures, some news arrives about the firm's future project payoffs. The new information is observed by both the

3. In costly state verification models the value of the borrower is not known until monitoring takes place Thus, even if the bank's junior claim is worthless, the bank does not know this until it monitors.

bank and the borrower, but it is not verifiable. Based on this information, and in particular if there is bad news, the borrower may choose to take a costly risk-increasing action. The contract may allow the bank to demand the collateral at this time (or, synonymously, the project liquidation value) instead of waiting for the contract to mature at date $t = 2$. Also at $t = 1$ the two parties may renegotiate the terms of the contract. Whether the borrower expends resources to add risk to the project, or whether instead the bank ends the contract early by seizing the collateral, depends on the outcome of renegotiation. Finally, if the project is not liquidated at $t = 1$, then at $t = 2$ the borrower repays the loan or is liquidated. If the borrower's project is not liquidated at $t = 2$, then a final payoff is received at $t = 3$. Figure 12.1 shows the timing of the model and Table 12.1 provides a concise summary of notation and definitions for future reference.

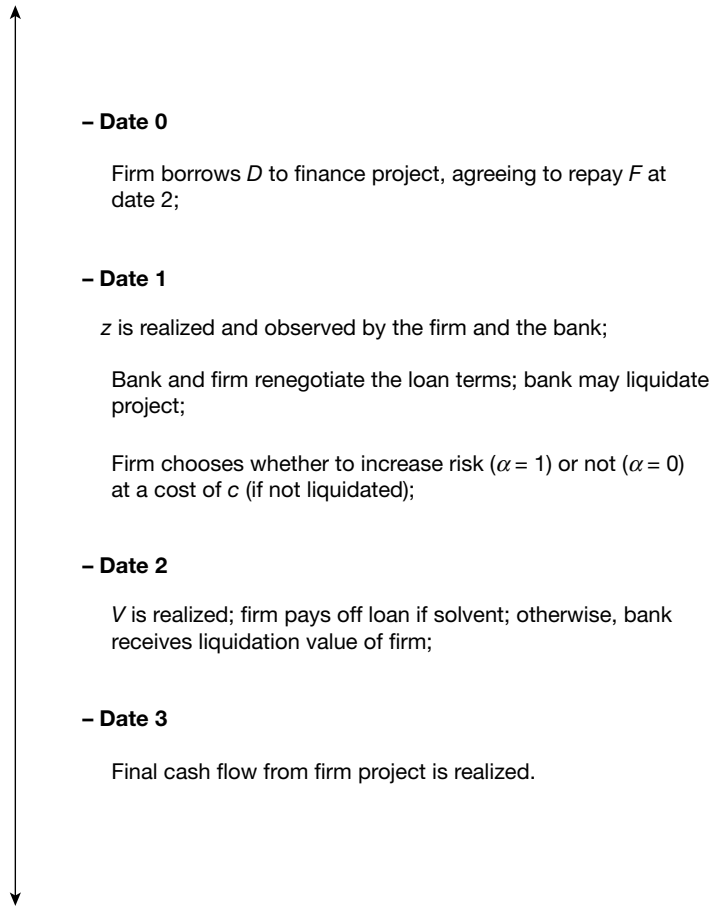


Figure 12.1 Sequence of events.

Table 12-1. SUMMARY OF SOME NOTATION

$y_2(z)$	Cash flow from the project at $t = 2$ for borrower type z
x_t	Scale of the project at time $t (t = 0, 1; x_0 = 1)$
D	Initial amount borrowed
F_0	Initial face value of the debt
V	Value of the project to the entrepreneur as of $t = 2$
L_t	Liquidation value of the project to the lender $(t = 1, 2)$
F^N	Renegotiated face value of debt at $t = 1$
α	Indicator variable for switching to riskier project, $\alpha \in \{0, 1\}$
c	Cost of switching to the riskier project
π^T	Total expected value of the project at $t = 1$
π^R	Expected bank profit with renegotiation
π^U	Expected bank profit absent renegotiation
z^*	inf $\{z \mid \Gamma(z) \geq 0$, given $F\}$, that is, threshold for switching to risky project given initial contract
z^{**}	inf $\{z \mid \Gamma(z) \geq F^N\}$, that is, threshold for switching projects given renegotiated contract
z_{RN}	inf $\{z \mid \pi^U \geq L_1$, for all $F^N \leq F\}$, threshold for liquidation to be a credible threat
z_{EL1}	inf $\{z \mid \pi^T \geq L_1, \alpha = 0\}$, threshold for efficient liquidation absent switching projects
z_{EL2}	Inf $\{z \mid \pi^T \geq L_1, \alpha = 1\}$, threshold for efficient liquidation given switching to the riskier project
z_{IL}	inf $\{z \mid \pi^R \geq L_1\}$, threshold for liquidation to be profit maximizing for the bank
$F^{++}(z)$	value of $F^N(z)$ that maximizes $\pi^R(F^{++}, z, \alpha = 1)$
$F^{-}(z)$	value of $F^N(z) > F$ that maximizes π^R given $\alpha = 1$

12.1.1. Detailed Assumptions of the Model

12.1.1.1. PROJECTS AND BORROWERS

The borrower's project requires a fixed scale of investment which, without loss of generality, we will set to one. The borrower has an amount $1 - D$ available to invest, but must obtain the remainder, D, externally.

The project generates cash flow realizations at dates $t = 2$ and $t = 3$ of $y_2(z)$, and $V(z)$, respectively, where z is the borrower type realized at $t = 1$. We refer to V as the project value, ignoring any liquidation possibilities (see below), and usually suppressing the dependence on z. For simplicity we assume a required rate of return of zero. The value V has a probability distribution given by $G(V; z, \alpha)$, where z, interpreted as "news" or borrower "type," is a random variable whose value is realized at $t = 1$, and where α indexes the project that the borrower selects at $t = 1$. (i.e., whether risk is added to the project). $G(V; z, \alpha)$,

is continuous and differentiable in V and z and has bounded support, $[V_l, V_h]$ We assume that:

Assumption 1. *Higher values of z represent "good news" in the sense that the conditional distribution of f (z|V) exhibits the monotone likelihood ratio property (MLRP), that is, f (z|V) /f (z|V*) is monotone in z, increasing if V > V*, and decreasing otherwise [see Milgrom (1981)].*

The random variable z, realized at $t = 1$, has density $h(z)$ and support $[z_l, z_h]$. We will refer to z as the borrower "type."

12.1.1.2. Liquidation Values

The project value as of $t = 2$, V, is to be interpreted as the net present value of the project when it is in the hands of the borrower who is assumed to have some special expertise relative to the bank. If the bank becomes the owner of the project, then it is worth a different value, the "liquidation" value or "collateral" value. Liquidation at date t means that the project yields L_t at that date in lieu of any future payoffs subsequent to the liquidation date. For simplicity, we assume:

Assumption 2. *Liquidation is all or nothing; liquidation values are certain and verifiable by both parties. Also,* $D > L_1 > L_2$.[4]

The last part of Assumption 2 says that the project requires outside financing in an amount that exceeds its liquidation value at any point in time so fully secured debt is not feasible.[5]

12.1.1.3. Asset Substitution by the Borrower

At $t = 1$ the borrower having received news, z, has the ability to unilaterally add risk to his project at a cost to the expected project return of c: adding risk reduces both V and L_2 by the amount c.[6] Adding risk, referred to as "asset substitution," is denoted by the discrete variable α (which equals 1 if the additional risk is taken and 0 otherwise).

Assumption 3. *Additional riskiness takes the form of a mean preserving spread:*

$$V_1 = V_0 + \in$$

4. The assumption that liquidation is all or nothing is without loss of generality since partial liquidation is never optimal in any case. We prove this in Gorton and Kahn (1992). Also, note that if $L_1 = L_2$, then the bank can never be worse off by allowing the project to continue at $t = 1$ and, thus, will never liquidate the project at that date. The assumption that $L_1 > L_2$ implies that at earlier stages of the project liquidation is less costly, that is, more can be recovered.

5. We also assume that there is no choice concerning collateral; the borrower uses all the collateral that the project provides and has no other collateralizable resources.

6. The assumption that the liquidation value, L_2, is also reduced by the amount c if risk is added ($\alpha = 1$) is not necessary, but appears (to us) to be realistic.

where V_α is the value of the project given choice α, and where $E(\in \mid V_0) = 0$.

We denote the distribution of $\in$ by $K(\in)$ and the density by $k(\in)$. The support of $\in$ is $[\in_l, \in_h]$.

Assumption 4. $V_0 + \in_l \geq c$.

Assumption 4 says that adding risk is always feasible; the borrower can always pay the cost c out of the project value when $\alpha = 1$ is chosen.

12.1.1.4. The Contracting Environment

The contracting environment is as follows:

Assumption 5. *The following are observable, but* not verifiable: *the borrower's project choice at $t = 1$, α; the project value, V; the realization of the borrower type, z; and the realized cash flow $y_2(z)$.*

Assumption 5 means that contracts can only be made contingent on the $t = 1$ liquidation value and payments by the borrower to the lender. These variables are observable by all parties, in particular, third-party contract enforcers.

12.1.1.5. Contracts and Renegotiation

A "bank" is distinguished from other providers of funds by:

Assumption 6. *Among possible funds providers, only banks can renegotiate at $t = 1$.*

According to Assumption 6, a bond blockholder who could carry out renegotiation is labeled a "bank" for our purposes.[7] Other fund providers are viewed as dispersed and incapable of coordinating renegotiating efforts. However, while we assume that an agent must be a bank in order to renegotiate, whether the contract includes the right to seize collateral prior to maturity is a separate issue.

Assumption 7. *A contract can include a provision allowing for the lender to seize the borrower's collateral at will at $t = 1$.*

We will call this contract provision the "liquidation option." Since the lender must decide when to seize the borrower's collateral, only banks would consider including this provision. This contract provision may be thought of as a reduced form for sufficiently detailed covenants that when violated allow the bank to

7. Thus the term "bank" is intended to apply to any agent who is the sole (or sufficiently large) lender to the borrowing firm and lends according to the contract we specify in the model. We do not intend the term to strictly apply to institutions chartered by the government, but rather to a broader class of agents, including so-called nonbank banks such as insurance companies, firms such as General Motors Acceptance Corporation, and agents who hold blocks.

demand collateral.[8] Exercising the liquidation option is infeasible for other creditors because, by assumption, other lenders cannot renegotiate and hence cannot initiate liquidation. Combining Assumption 6 and Assumption 7 means that there are three distinct securities to consider: corporate bonds (dispersed holders who cannot renegotiate), and bank loan contracts with and without the liquidation option.

In order to most simply characterize the renegotiation outcomes at $t = 1$, we assume that:

Assumption 8. *The bank can credibly make a take-it-or-leave-it offer at* $t = 1$.

Assumption 9. *Borrowers have no alternative source of financing at the date of renegotiation,* $t = 1$.

The outcome of renegotiation at $t = 1$ will either be liquidation of the project or a contract specifying a payment to be made at $t = 2$ (either on new terms or at the status quo ante). Because cash flows are not verifiable, they can be consumed by the borrower; they cannot be seized by outside lenders, such as the bank, but may be handed over voluntarily by the borrower. In this setting Kahn (1992) shows that debt is an optimal contract.[9] For the purposes of this article we assume that:

Assumption 10. *Debt is the optimal contract from* $t = 1$ *to* $t = 2$. *Failure to repay the debt at* $t = 2$ *triggers liquidation, that is, the parties are committed to liquidation if there is a default.*

In order to avoid liquidity problems, we assume that the cash flow at $t = 2$, $y_2(z)$, is sufficiently high, for all z, so that it is feasible to repay the lender at $t = 2$ if the borrower so chooses.

12.1.1.6. Opportunism by the Bank

When the bank has the opportunity to threaten liquidation early (because this contract provision has been included) it may use this threat to simply extract surplus from the borrower. We will call this "opportunism." Bank opportunism will sometimes have efficiency considerations. Let $\pi^R(F^N, z, \alpha)$ be the expected profits of the bank as of $t = 1$ after renegotiation has resulted in a new face value for the debt of F^N. (α is a function of F^N and z, but for clarity we include it as an argument of the expected profit function.) F^N could be higher or

8. In the United States, bank loan contracts contain detailed covenants which are easily violated, triggering the bank's right to demand collateral even if the borrower has not missed a payment on the loan. In other countries, such as Japan, the loan contract is more straightforward in stating that the bank has the right to demand collateral any time.

9. Bolton and Scharfstein (1990) and Hart and Moore (1998) show the optimality of debt in similar settings.

lower than the initial face value. If the bank can succeed in obtaining a higher rate, it faces a choice: raise the rate to maximize expected profit, accepting that the borrower will choose $\alpha = 1$ (call this rate F^{++}); or raise the rate to the highest level so that the borrower just chooses $\alpha = 0$ (call this rate F^{+}).

Assumption 11. $\pi^R(F^{++}, z, \alpha = 1) > \pi^R(F^{+}, z, \alpha = 0)$, *for all z.*

Assumption 11 means that bank opportunism has efficiency considerations since, if it can, the bank will renegotiate an interest rate which is so high that the borrower will add risk, even if the borrower would not add risk at the initial interest rate. Assumption 11 is not the only case in which bank opportunism will have efficiency considerations. Furthermore, it is not necessary for the analysis, but it is the most interesting case. The alternative assumptions are discussed further below and results for these cases are given in Appendix C.

12.1.1.7. Parameter Restrictions

Appendix A details three further assumptions concerning parameters of the model. Assumption 12 ensures that adding risk always results in a positive probability of solvency. This assumption simply makes the problem interesting since it says that when risk is added there is always some chance for the borrower to benefit. Assumption 13 guarantees that the bank always prefers that the borrower not add risk. Again, this is the interesting problem since otherwise the bank would not want to prevent asset substitution. Finally, Assumption 14 ensures that bank profits are increasing in the economically relevant range of *F*. The assumption allows us to ignore this issue of debt forgiveness (which has no efficiency considerations).

12.1.2. Discussion of the Model

Renegotiation occurs when news, *z*, arrives and is observed by both parties to the contract. Bank loans include covenants which require the firm to supply regular accounting information and provide the bank with an opportunity to investigate the firm.[10] Thus we view it as reasonable that the bank can observe *z,* which should be interpreted as new information about the firm's prospects that is not freely available to (or easily interpretable by) the public.

The timing of the model assumes that news (z) arrives before the cash flows. This is for simplicity. Since the loan cannot mature before sufficient cash flows from the project are realized, there is always potential for renegotiation during

10. Zimmerman (1975), Quill, Cresci, and Shuter (1977), and Morsman (1986) describe real-world covenants. Rajan and Winton (1995) discuss the theoretical rationale for their existence.

the course of the loan. We simply label the arrival of news and the consequent renegotiation as $t = 1$, but in principle these events can occur at any time prior to maturity, provided the borrower has time to add risk if he so chooses.

Renegotiation at $t = 1$ is complicated by two moral hazard problems. The first moral hazard problem concerns the borrower. The borrower can threaten to add risk to the project in order to transfer value from the bank. Adding risk is costly because it reduces the project value, V, and the liquidation value, L_2, by c. This can be interpreted as a transaction cost; the borrower must pay to modify the existing project so as to increase riskiness.[11] We will show below that our assumptions restrict attention to cases where the added risk is inefficient. Obviously if the additional risk is in the interest of both parties, then such an action should, and will, be taken and we do not concern ourselves with it (by Assumption 13).

The other moral hazard problem is bank opportunism. The bank may opportunistically threaten to liquidate in order to extract surplus from the borrower once news, z, has arrived. If the bank has the power to threaten liquidation and can thereby extract surplus from the borrower, it may behave inefficiently. Indeed, Assumption 11 says that the bank will behave this way if it has a credible liquidation threat. Of importance, this opportunism has efficiency considerations since the borrower will choose to add risk ($\alpha = 1$) when the bank behaves opportunistically.

The credibility of this threat by the bank depends on the design of the contract. The contract design problem involves the considerations discussed in the introduction. First, is renegotiation desirable? If it is, then should the contract with the bank include a provision that allows the bank to ask for the collateral prior to maturity of the loan? We assume that the contract can feasibly include the liquidation option which allows the bank to "call the loan" at $t = 1$ if it so wishes, and we ask whether it is optimal to include this provision.[12] If the liquidation option is included, then the third contract design consideration involves the specification of the initial ($t = 0$) contract form. Knowing that any contract

11. At $t = 1$ we assume that costless, or extremely inexpensive, ways of adding risk can be prevented costlessly by the bank through covenant restrictions.

12. The interpretation of this is that while borrower type, z, is not verifiable, a contract can contain verifiable provisions (covenants) which are always triggered by the arrival of the news, z. Loan covenants are written in terms of variables measurable according to accounting procedures, for examples, net worth, leverage, etc., and consequently are verifiable, though violations may be forgiven by the bank. See Zimmerman (1975), Quill, Cresci, and Shuter (1977), and Morsman (1986). Bank loan contracts are written with a large number of covenants so that small deviations of the state of the firm trigger covenant violations, allowing the firm to "call" the loan. Sometimes the bank excuses such violations. Because of these covenants, the option to "call" is best viewed as always verifiably being "in the money" for bad borrowers.

will be renegotiated at $t = 1$, what contract should be signed at $t = 0$? Our analysis attacks this question by asking: What is the gain to specifying the face value of the debt to be paid at $t = 2$, denoted F_0, at $t = 0$? In our analysis it is feasible for the parties to specify $F_0 = D$ at $t = 0$ (or, for that matter, $F_0 = \infty$). For example, specifying $F_0 = D$ would be tantamount to an initial agreement under which the lender essentially says to the borrower: "Here's an amount of money, D. I have the right to liquidate at $t = 1$, at which time we'll work out the details of the contract." This specification of the initial contract says that the bank can threaten to liquidate all borrower types at $t = 1$, receiving L_1, unless borrowers agree to the bank's offer of F^N at that date. We will show how renegotiation outcomes are affected by the specification of F_0 at $t = 0$ even though it is common knowledge that renegotiation will occur. The range of borrower types for which the liquidation threat is credible depends on the initial specification of F_0. The size of F_0 will lead to efficiency considerations via its ability to influence the bank's bargaining power at $t = 1$. The costs and benefits of allocating power to the bank will determine the initial F_0.

Since we have assumed that the borrower has no alternative financing source at $t = 1$, the borrower cannot threaten to refinance from other sources. It will also turn out that the bank's ability to threaten the borrower is limited. Thus it is not obvious how the surplus at $t = 1$ will be split. We have assumed that the bank can credibly make a take-it-or-leave-it offer at $t = 1$ and hence can obtain all the surplus. Since banking is competitive at $t = 0$, the possibility of extracting surplus at date $t = 1$ will be priced ex ante. The surplus will be split differently if other bargaining games are allowed, but this will not effect our results concerning efficiency.

12.2. DEFINITIONS AND PRELIMINARY LEMMAS

In this section we provide preliminary definitions and results. We prove two lemmas to build understanding of the model. First, we analyze the borrower's decision at date $t = 1$ concerning adding risk. This defines a critical borrower type z^* below which the borrower will add risk in the absence of any bank action. Then we show that adding risk is inefficient. We then define the payoffs relevant to the subsequent analysis. Finally, we outline the possible renegotiation outcomes and provide some intuition before the formal analysis.

12.2.1. News Arrival, the Borrower's Project Choice at $t = 1$, and Efficiency

At $t = 1$ the borrower and lender observe the realization of borrower type, z. The realization of a low z means that the borrower's equity is worth less than it was

ex ante. In this situation, as is well known, the borrower may have an incentive to switch projects to add risk ("asset substitution"). Borrowers who receive bad news (low z realizations) will be tempted to switch from their initial project, $\alpha = 0$ to a higher risk project, $\alpha = 1$. By increasing the variance of the project, the value of the firm's equity can be increased at the expense of the bank. But since it is costly to take this action, only firms with sufficiently bad "news" will choose $\alpha = 1$, as the following lemma shows.

LEMMA 1. *Given F_0, there exists some z^* such that setting $\alpha = 1$ is profitable for the borrower if and only if $z < z^*$. Furthermore, z^* is increasing in F_0.*

Proof. See Appendix B.

The lemma establishes that there is a critical borrower type, z^*, below which borrowers choose to add risk to their projects. Define the gain to the bank from the borrower of type z adding risk to be $\Gamma_B(z;\ F)$; see Appendix B. Then z^* is defined by $\Gamma(z^*;\ F_0) = 0$. We refer to $z < z^*$ as "bad" borrowers, and to $z > z^*$ as "good" borrowers. Also, eventually, we solve for F_0, the initial face value of the debt. In this regard, it is important to know how z^* depends on F_0, since lenders will take adverse incentive affects of higher F_0 into account initially and during any renegotiation. As the lemma shows, the dependence is intuitive: the higher the borrower's debt burden, the more likely it is that asset substitution will be appealing.

Lemma 1 shows that borrowers of type $z < z^*$ will, ceteris paribus, add risk. Our focus is on situations where the risk taking by the borrower is unprofitable for the bank and socially inefficient. The next lemma shows that, under our assumptions, this is ensured.

LEMMA 2. *The addition of risk by the borrower ($\alpha = 1$) is unprofitable for the bank.*

Proof. See Appendix B.

It follows immediately that since asset substitution by the borrower is always bad for the bank, it is socially harmful on the margin. That is, for some range of $z < z^*$, a borrower of type z^* is indifferent to adding risk while the bank strictly prefers that risk not be added. Figure 12.2 depicts typical "gain" functions for the borrower and lender. Lemmas 1 and 2 only say that the gain for the borrower crosses zero somewhere from above, while the gain for the lender is always negative under our assumptions. Thus the sum of the two gains (which represents the net social gain from asset substitution) will cross zero to the left of z^*. This implies that there is a range of z values to the left of z^* such that asset substitution is inefficient yet is in the private interest of the borrower absent preemptive action by the bank.

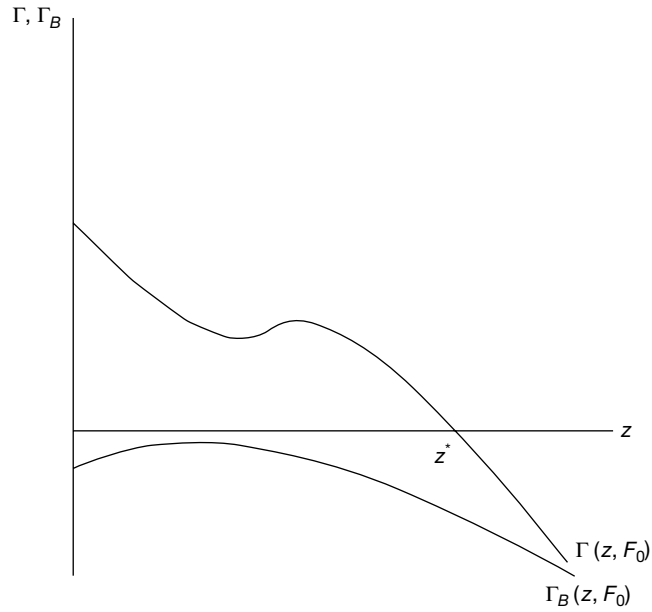


Figure 12.2

12.2.2. Payoffs

At $t = 1$ the bank may liquidate the project or renegotiate the interest rate. Let F^N be the new (i.e., renegotiated) face value for the debt to be paid at $t = 2$. In general, F^N will depend on z, but this notation is usually suppressed.

Define the total expected payoff to the project as of $t = 1$ for given z and choice of α, $\pi^T(F^N, z, \alpha)$, as follows:

$$\pi^T(F^N, z, \alpha) \equiv (L_2 G(F^N(z) \,|z, \alpha) + \int_{F^N(z)}^{V_h} Vg(V \,|z, \alpha)\, dV + y_2(z) - \alpha c. \tag{12.1}$$

Note that this is not the first-best total expected value, but the second best.[13] Define unrenegotiated bank profit, $\pi^U(F_0, z, \alpha)$, to be expected bank profit as of $t = 1$, from a borrower of type z, when evaluated at the initial face value of the debt, F_0, given that the borrower chooses α according to whether $z < z^*$:

$$\pi^U(F_0, z, \alpha) \equiv (L_2 - \alpha c)G(F_0|z, \alpha) + F_0[1 - G(F_0|z, \alpha)], \tag{12.2}$$

where α is a function of F_0 and z.

13. The payoff is second best because sometimes L_2 is obtained due to default on the debt. Under first best this would not happen.

To facilitate discussion of liquidation define:

$$z_{\mathrm{EL1}} = \inf\{z : \pi^T(F^N, z, \alpha = 0)L_1\};$$
$$z_{\mathrm{EL2}} = \inf\{z : \pi^T(F^N, z, \alpha = 1)L_1\};$$
$$z_{\mathrm{IL}} = \inf\{z : \max[\pi^R(F^N, z, \alpha = 1), \pi^U(F_0, z, \alpha = 1)] = L_1\}.$$

The point z_{IL} is defined as the lowest borrower type at which the best the bank can do under any renegotiation strategy (including not renegotiating) is just equal to the liquidation value of the project. As will become clear, the subscript "EL1" denotes first-best efficient liquidation because the value of projects of type lower than z_{EL1} is expected to be less than the liquidation value of the project even if the borrower does not add risk. The subscript "EL2" denotes second-best efficient liquidation, indicating that the value of projects of type $z_{\mathrm{EL1}} < z_{\mathrm{EL2}}$ is expected to be less than the liquidation value *only* if the borrower chooses to add risk. If the borrower does not add risk, then these projects should not be liquidated (from the point of view of a social planner). Note that $z_{\mathrm{EL1}} < z_{\mathrm{EL2}}$. The reason for this inequality is that switching to $\alpha = 1$ reduces the expected return because it costs c to switch projects. The subscript "IL" denotes inefficient or excessive liquidation because, as will be seen, some projects of type $z < z_{\mathrm{EL2}}$ may be liquidated. z_{IL} is defined with respect to the bank's expected profit and thus will define when liquidation occurs. Consequently, z_{IL} may or may not coincide with z_{EL2}, as seen below.

12.2.3. Renegotiated Interest Rates

If the bank does not liquidate the borrower's project, it may seek to renegotiate the interest rate on the loan.[14] In this subsection we outline the possible renegotiation outcomes (to be analyzed subsequently) and provide some intuitive explanation. The intuition follows the ordering of the z-cutoff points shown in figure 12.3.

Define renegotiated bank profits at $t = 1$, when a new interest rate $F^N(z)$ has been agreed to as follows:

$$\pi^R(F^N, z, \alpha) \equiv (L_2 - \alpha c)G(F^N(z) \mid z, \alpha) + F^N(z)[1 - G(F^N(z) \mid z, \alpha)]. \quad (12.3)$$

Again, α is the same function of F and z. Renegotiated bank profit is the return the bank expects to receive from the project of a borrower of type z, where the

14. In fact, even absent the moral hazard problem of asset substitution, it would be in the bank's interest to change F upon learning z simply to increase expected payoffs. We postpone discussion of this until later.

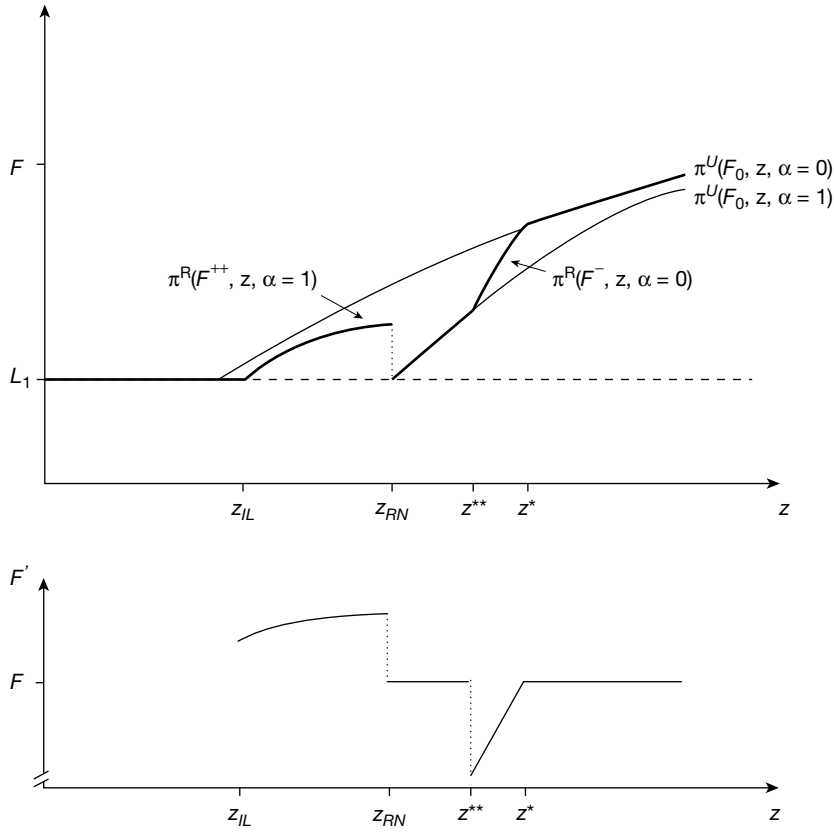


Figure 12.3

borrower of type z chooses project α, and promises to repay $F^N(z)$, the new interest rate agreed upon at date $t = 1$.

One possible renegotiation outcome would be a *lower interest rate*. For example, if the borrower type is such that the gain to switching projects is positive, that is, $z < z^*$, then the bank may forgive part of the debt by lowering the interest rate to induce the borrower not to add risk (switching to $\alpha = 1$). Consider a borrower of type just worse (i.e., lower) than z^*. Such a borrower will choose to add risk, $\alpha = 1$, but is near indifference. If the value of the borrower's equity were a little higher, then $\alpha = 0$ would be chosen so the cost c would not be borne. The bank may find it profitable to raise the value of the borrower's equity by forgiving some debt. While this lowers the face value of what the borrower contracts to repay, the bank's expected profits may rise because the borrower, with reduced leverage, chooses not to add risk. (In fact, it is possible that the bank would want to forgive debt even for some $z > z^*$, simply because it improves expected profit.) In any case, define $F^-(z)$ to be the highest value of F such that

$\alpha = 0$ solves the borrower's $t = 1$ problem of maximizing the (expected) gain to adding risk.

It need not be the case that $F^{-}(z) < F_0$. But if $F^{-}(z^*) < F_0$, then for some range of borrowers in the interval $z_1 < z < z^*$, the bank may want to forgive debt. But at some point, for sufficiently low z, lowering the interest rate to induce the borrower not to switch projects will reduce the bank's expected profit below what it would earn if it maintained the initial contract (F_0) and allowed the borrower to add risk $(\alpha = 1)$. Define z^{**} to be the borrower type at which the bank is indifferent between these two choices: $\pi^{R}(F^{-}, z = z^{**}, \alpha = 0) = \pi^{U}(F_0, z = z^{**}, \alpha = 1)$, where $\pi^{R}(F^{-})$ is the bank's expected profit as of $t = 1$ when the renegotiated interest rate is decreased $[\pi^{R}(F^{+})$ will indicate expected bank profit when the renegotiated interest rate is increased]. Note that by definition it is always the case that $z^{**} < z^*$; the borrower would only be tempted to choose $\alpha = 1$ if $z < z^*$, that is, when the gain to switching projects is positive $(\Gamma(z) > 0)$. Thus z^{**} is the threshold value of z below which (even with renegotiation) the borrower chooses $\alpha = 1$. See figure 12.3.

Since z^{**} defines the point at which borrowers add risk, it will be important to know how this point varies with F_0. The answer is given by:

LEMMA 3. *z^{**} is increasing in F_0.*

Proof. Note that $\pi^{R}(F^{N}, z, \alpha)$ is independent of F_0, but $\partial \pi^{U}/\partial F_0 > 0$ for $F_0 < F^{\#}$, by Assumption 13. Since z^{**} is defined as the point where $\pi^{R}(F^{-}, z = z^{**}, \alpha = 0) = \pi^{U}(F_0, z = z^{**}, \alpha = 1)$ the lemma follows. □

If forgiving debt to induce the borrower to choose $\alpha = 0$ is not profitable, then the bank may seek to raise the interest rate, provided it has a credible (i.e., subgame perfect) threat to liquidate. Define z_{RN} to be the solution to $\max[\pi^{U}(F_0, z_{RN}, \alpha), \pi^{R}(F^{-}, z_{RN}, \alpha)] = L_1$ and if $\pi^{U} > L_1$, for all z, then $z_{RN} = z_1$. For $z < z_{RN}$ the bank expects its (unrenegotiated) profit to be less than the current liquidation value and hence has a credible threat to liquidate. The subscript "RN" denotes renegotiation since for $z < z_{RN}$ the bank can credibly threaten the borrower and demand a higher interest rate. If the bank can credibly threaten the borrower, then the higher interest rate is given by:

$$F^{++}(z) = \text{Argmax}_{F^N}\,(L_2 - \alpha c)\,G\left(F^{N}|z, \alpha\right) + F^{N}\left[1 - G\left(F^{N}|z, \alpha\right)\right]. \tag{12.4}$$

Recall that under Assumption 11, the bank's expected profit is higher if it raises the interest rate so much that the borrower adds risk, as opposed to raising it to $F^{+}(z)$ and receiving $\pi^{R}(F^{+}(z), z, \alpha = 0)$.

As shown in figure 12.3, as the type of the borrower declines, there comes a point where raising the interest rate cannot raise the expected value of the

loan to the bank above the liquidation value, L_1. As defined above, at z_{IL}, $\pi^R(F^{++}, z_{IL}, \alpha = 1) = L_1$. Again, however, it is good to keep in mind that there can be other cases where the bank can profitably raise the interest rate.

As with the other critical z-values, z_{RN} depends on F_0.

LEMMA 4. *z_{RN} is decreasing in F_0.*

Proof. When $F_0 < F^{\#}$, $\partial \pi^U / \partial F_0 > 0$, by Assumption 13. □

12.2.4. Definition of Equilibrium at $t = 1$ and Specification of Cases

At $t = 1$ the bank and the borrower know L_1, observe the realization of z, and choose a new contract, F^N, or liquidation, subject to constraints imposed by the existing contract, F_0. The existing contract and the borrower's type determine $\pi^U(F_0, z, \alpha)$, that is, the unrenegotiated expected bank profit. An equilibrium at $t = 1$ is (1) a choice of α by a z-type borrower which maximizes the borrower's expected profits, given the new contract, $F^N(z)$ (assuming liquidation does not occur); and (2) a choice of (new) interest rate, $F^N(z)$, or liquidation, by the bank, given the borrower's type, z, and choice of α, which maximizes the bank's expected profit. The resulting bank profit function, which we will denote by $\pi^B(F^N, z, \alpha)$, is the upper envelope of the four profit functions based on the different renegotiation outcomes, that is, $\pi^B(F^N, z) = \max\{\pi^R(F^{++}, z, \alpha), \pi^R(F^-, z, \alpha), \pi^U(F_0, z, \alpha), L_1\}$, given the optimal choice of risk, α, by the borrower as a function of z.

The precise pattern of renegotiation outcomes as a function of z depends on the location of z_{RN} relative to z^{**} and z^*. These in turn depend on F_0 and L_1. We treat L_1 as fixed and let F_0 trace out all of the possibilities, although of course ultimately F_0 will be determined by equilibrium conditions. Lemmas 1, 3, and 4 imply that there are three scenarios to consider (as depicted in figure 12.4) corresponding to low, intermediate, and high values of F_0.[15] At low values of F (the bottom panel of figure 12.4 the bank has a credible threat to renegotiate over a wide range of z, so $z_{\mathrm{RN}} > z^*$ and there is never any issue of forgiving debt. The bank just "holds up" everyone with $z \leq z_{RN}$, even knowing that they will add risk as a consequence.

At intermediate and high values of F_0 we have $z_{\mathrm{RN}} > z^*$, so there is a range of forgiveness. The difference between the two is that with high F_0, $z_{\mathrm{RN}} < z^{**}$ so there is a range in which risk taking occurs because it is not in the bank's interest to forgive. The loan is still profitable though, so the bank has no credible threat

15. In figure 12.4 the curve labeled $\pi^R(F^-)$ is $\pi^R(F^+)$ for $z < z^*$ and $\pi^R(F^-)$ for $z < z^*$. To avoid complicating the figure we only include one label.

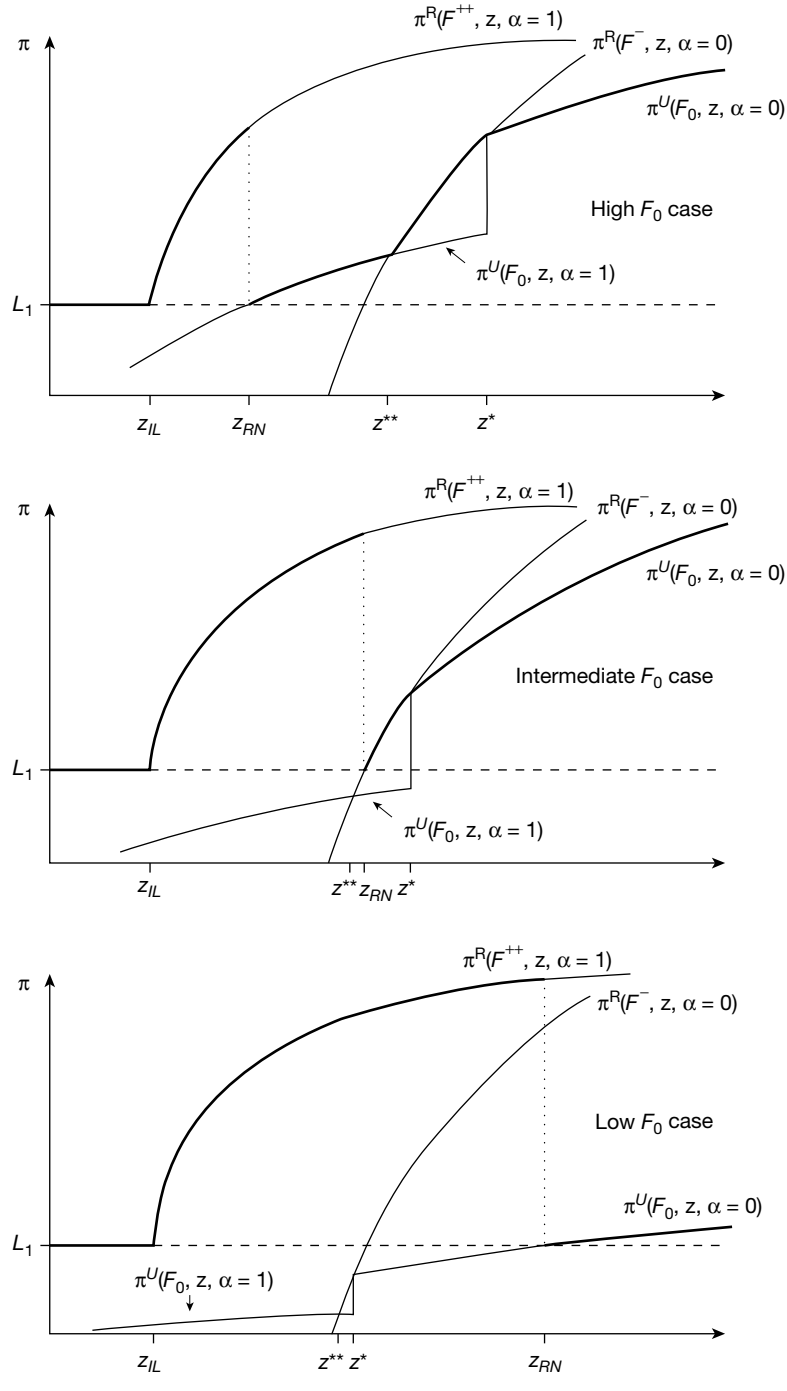

π
$\pi^R(F^{++}, z, \alpha = 1)$
$\pi^R(F^-, z, \alpha = 0)$
$\pi^U(F_0, z, \alpha = 0)$
High F_0 case
$\pi^U(F_0, z, \alpha = 1)$
L_1
z_{IL}
z_{RN}
z^{**}
z^*
π
$\pi^R(F^{++}, z, \alpha = 1)$
$\pi^R(F^-, z, \alpha = 0)$
$\pi^U(F_0, z, \alpha = 0)$
Intermediate F_0 case
L_1
$\pi^U(F_0, z, \alpha = 1)$
z_{IL}
z^{**} z_{RN} z^*
π
$\pi^R(F^{++}, z, \alpha = 1)$
$\pi^R(F^-, z, \alpha = 0)$
Low F_0 case
$\pi^U(F_0, z, \alpha = 0)$
L_1
$\pi^U(F_0, z, \alpha = 1)$
z_{IL}
z^{**} z^*
z_{RN}
Indicates $\pi^B(F, z, \alpha) \equiv \max \{\pi^R(F^{++}, z, \alpha), \pi^R(F^-, z, \alpha), \pi^U(F_0, z, \alpha), L_1\}$.

Figure 12.4

that would allow it to increase F either, and it just leaves it at F_0. In the intermediate case the forgiveness range runs into the "hold-up" range, so risk taking coincides with the bank's increasing F.

It will turn out that the equilibrium value of F_0 corresponds to the boundary between the intermediate and high F_0 cases, with $z^{**} = z_{RN}$. This is because, as figure 12.4 makes clear, the range of risk taking (which occurs for $z < \max[z^{**}, z_{RN}]$ is thereby minimized. In the next two sections we will go into more detail on the high F_0 case, and relegate the other cases to Appendix C.

12.3. RESULTS: RENEGOTIATION AND LIQUIDATION DECISIONS AT T=1

Having dispensed with the preliminaries, we can now turn to the actual predictions of the model. The real point of interest in the model is at $t = 1$, when all the important decisions get made. At that point project-specific information has arrived, and the borrower and lender have to decide whether to continue the project and if so, on what terms. At $t = 2$, behavior is mechanical—the borrower repays the loan at whatever the prevailing terms are if the project is solvent, or he does not, and the project is liquidated. At $t = 0$ all borrowers are identical, so the only problem is to determine the initial face value of the loan, a problem we turn to in Section 12.4.

12.3.1. The Liquidation Decision

What triggers liquidation? By definition of z_{IL}, projects of borrowers of type $z < z_{IL}$ are liquidated. In the high F_0 case, liquidation begins at the point where $\pi^R(F^+, z_{IL}, \alpha = 1) = L_1$. If $\pi^T(z_{IL}, \alpha = 1) = \pi^R(F^{++}, z_{IL}, \alpha = 1) = L_1$, (i.e., $z_{EL2} = z_{IL}$), then the projects liquidated in the range $z_{EL1} < z < z_{IL}$ are second-best liquidated since total expected profits are positive if the borrower did not choose $\alpha = 1$. However, if $\pi^T(z_{IL}, \alpha = 1) > \pi^R(F^+, z_{IL}, \alpha = 1) = L_1$, then $z_{IL} > z_{EL2}$, and even more projects are liquidated, inefficient (or excessive) liquidation ("IL") beyond the second best. This inefficient liquidation (relative to second best) can happen because there is no way for the bank to overcome the incentive the borrower has to choose more risk. Forgiveness does not increase the bank's expected profit by enough, nor does raising the interest rate. (We discuss the issue of side payments below.)

Liquidation of socially wasteful projects will be an important role for the bank to play. But by giving the bank the power to liquidate there is also the possibility

that the bank liquidates projects inefficiently. This cost will have to be weighed against the benefits of liquidating efficiently.

12.3.2. Renegotiation Outcomes

We now turn to renegotiation with borrowers who are not liquidated, maintaining the focus on the high F_0 case. Renegotiation outcomes, as a function of borrower type, are characterized by the bank choosing the outer envelope of four expected profit curves: renegotiated profit when the interest rate is raised, $\pi^R\left(F^{++}, z, \alpha = 1\right)$; renegotiated profit when debt is forgiven (i.e., the interest rate is lowered), $\pi^R\left(F^{-}, z, \alpha = 0\right)$; unrenegotiated profit, $\pi^U(F_0, z, \alpha)$; and liquidation. Figure 12.3 graphically portrays the four bank profit curves in the high F_0 case. The next proposition formalizes the intuition that the bank will choose the outer envelope of these profit curves subject to its ability to extract surplus from the borrowers.

PROPOSITION 1. *In the high F_0 case, renegotiation results in*

(*i*) $F^N(z) = F_0$ *for all* $z > z^*$, *that is, no change in the interest rate. The borrower chooses* $\alpha = 0$.
(*ii*) $F^N(z) = F^-(z) < F_0$ *for all* $z \in [z^{**}, z^*]$, *that is, forgive debt (lower the rate) so that the borrower chooses* $\alpha = 0$.
(*iii*) $F^N(z) = F_0$ *for all* $z \in [z_{\mathrm{RN}}, z^{**}]$, *that is, no change in the interest rate. The borrower chooses* $\alpha = 1$.
(*iv*) $F^N(z) = F^{++}(z) > F_0$ *for all* $z \in [z_{\mathrm{IL}}, z_{\mathrm{RN}}]$, *that is, raise the interest rate and let the borrower choose* $\alpha = 1$.

Proof. See Appendix B.

Intuitively the proposition says the following: Upon arrival of news at $t = 1$, there are four potential outcomes in addition to immediate liquidation:

1. With favorable news, the status quo obtains, as the borrower is not interested in asset substitution and the bank has no credible threat to liquidate the project and thereby extract a higher interest rate through renegotiation.
2. With moderately unfavorable news, the bank will choose to forgive some of the debt (i.e., lower the interest rate) in order to induce the borrower not to engage in costly asset substitution.
3. With more unfavorable news, however, the bank will not be able to preclude asset substitution by offering debt forgiveness. Instead, the asset substitution will occur and the project will become more risky.
4. Finally, with the most unfavorable news, asset substitution will occur but the bank will be able to extract a higher interest rate through

renegotiation because the project's prospects are so poor that the bank has a credible threat to liquidate.

Thus the bank is unable always to preclude asset substitution and the resulting endogenous increase in project risk. It will turn out that in equilibrium cases 3 and 4 above coincide; that is, the bank will either forgive some of the debt to preempt asset substitution, or it will concede the substitution and extract a higher interest rate. The status quo is never the best option once bad news arrives.

The proposition can also be understood with reference to figure 12.3. Starting with the highest type borrowers, those with $z > z^*$ unrenegotiated bank profits are given by $\pi^U(F_0, z, \alpha = 0)$ since these borrowers do not switch projects. The bank cannot credibly threaten these borrowers to extract a higher rate because in this range, $\pi^U(F_0, z, \alpha = 0) > L_1$ (that is, $z_{RN} < z^*$). The bank may or may not forgive debt for these borrower types (we assume that there is no forgiveness by Assumption 13), but in any case these borrowers choose $\alpha = 0$. Therefore these borrowers continue their projects and the bank maintains the initial interest rate F_0. This is shown in the lower panel of the figure.

Borrowers with types below z^* will choose to add risk to their projects, ceteris paribus. But the bank is not in a position to threaten all of these borrowers with liquidation because the point at which the bank can credibly threaten and force renegotiation, z_{RN}, is below $z^*(z_{RN} < z^*)$. However, by providing debt forgiveness to some of these borrowers they can be induced to not add risk. Debt forgiveness raises the value of the borrower's equity by just enough to make taking the costly, risk-increasing, action unprofitable. The question is whether this is profitable for the bank. In the figure it can be seen that the bank's expected profit when debt is forgiven (that is, the interest rate is lowered to $F^-(z) < F_0$) is higher than unrenegotiated bank profits given that borrowers choose $\alpha = 1$. (The interval $[z^{**}, z^*]$ may not exist.)

Debt forgiveness is optimal as long as $\pi^R(F^-, z, \alpha = 0) > \pi^U(F_0, z, \alpha = 1)$, that is, until the bank must forgive so much debt that it prefers to stay with the initial contract and allow the borrower to add risk. At the point z^{**}, $\pi^R(F^-, z^{**}, \alpha = 0) = \pi^U(F_0, z^{**}, \alpha = 1)$, so debt forgiveness is only provided for borrowers of type $z^{**} < z < z^*$ since they can be induced to not add risk, which is in the bank's best interest. For borrowers in the range $z_{RN} < z < z^{**}$ there is no change in the interest rate since these borrowers cannot be threatened to get a higher rate and debt forgiveness is not profitable. Consequently, borrowers of type $z_{RN} < z < z^{**}$ are allowed to add risk and continue under the old contract. This is shown in the bottom panel of the figure where these borrowers continue with an interest rate of F_0.

For borrowers of type $z_{IL} < z < z_{RN}$ it is not profitable for the bank to forgive debt (since $z^{**} > z_{RN}$), but the project is worth continuing. The bank can force the borrower to pay a higher interest rate because the threat of liquidation is

credible for these borrower types [since $\pi^U(F_0, z, \alpha = 1) < L_1$ in this range]. Finally, at z_{IL} $\pi^R(F^{++}, z_{IL}, \alpha = 1) = L_1$, so borrowers of lower type than this are liquidated.

Proposition 1 covers the case assumed by Assumption 11, that it is always more profitable for the bank to raise the rate to F^{++} and let the borrower add risk, if the bank can credibly threaten liquidation. Appendix B analyzes the alternatives to Assumption 11 as well as the high F_0 and low F_0 cases.

12.3.3. Discussion

Two features of Proposition 1 are worth noting. First, the bank is not entirely successful in controlling risk. Borrowers of type $z_{IL} < z < z^{**}$ choose to add risk and are allowed to continue their projects. Thus, in equilibrium, borrower risk varies endogenously. Second, renegotiated interest rates are not monotonic in borrower type as can be seen in the lower panel of figure 12.3. Starting from z^*, the bank first lowers the interest rate to forgive debt (until z^{**} is reached), then maintains the initial rate (until z_{RN} is reached), and then raises the rate (until z_{IL} is reached) after which projects are liquidated.

We have allowed for the possibility that the bank may increase F if it has a credible threat to liquidate, regardless of whether the borrower will choose to add risk or not. We have postponed until now the possibility of debt forgiveness simply as the result of new information being received at $t = 1$, namely z. Even absent any moral hazard problem, the bank may be able to increase its expected profits by lowering F for some borrowers. This possibility would only change the shape of the π^U functions monotonically without qualitatively changing figure 12.3 or any of the results described above. In particular, without the moral hazard problem, these reoptimized interest rates would introduce no new nonmonotonicity in the pattern of renegotiated interest rates as a function of borrower type z.

12.4. INITIAL LOAN PRICING AND THE ROLE OF DEBT

The renegotiation outcomes at $t = 1$ were determined above assuming that the contract contained the liquidation option and assuming a given F_0 that had been determined earlier at $t = 0$. If the liquidation option is not included in the contract, then the bank, being a single agent, can renegotiate, but cannot threaten liquidation. Before considering the optimality of the liquidation option, which is done in Section 12.6, we turn to the determination of F_0 in the case where the liquidation option is included in the contract. In this case, both parties to the contract know that renegotiation can occur. Then, what role does F_0 play? Why bother specifying F_0, at all, given that it is renegotiated after news arrives?

To answer these questions we proceed in two steps. First, we demonstrate how efficiency considerations determine F_0 by affecting the bargaining power of the bank. This will determine the F_0 that is socially optimal (in the second-best sense). Then we inquire as to how the (second-best) efficient F_0 can be implemented when lenders act competitively and earn zero expected profits.

12.4.1. The Socially Optimal Initial Interest Rate

The socially optimal (second-best) F_0, call it F_0^* will minimize inefficient risk-taking subject to the moral hazards. To determine F_0^* we first need to decide which of the three cases defined above, high F_0, low F_0, or intermediate F_0, is most efficient. We can summarize the analysis so far, with respect to which borrowers will add risk to their projects, by combining the results of Proposition 1 with the results in Appendix B:

Low F_0 case: For $z_{\text{IL}} < z < z_{\text{RN}}$, $\alpha = 1$, while for $z_{\text{RN}} \leq z \leq z_h$, $\alpha = 0$.

Intermediate F_0 case: For $z_{\text{IL}} < z < z_{\text{RN}}$, $\alpha = 1$, while for $z_{\text{RN}} \leq z \leq z_h$, $\alpha = 0$.

High F_0 case: For $z_{\text{IL}} < z < z^{**}$, $\alpha = 1$, while for $z^{**} \leq z \leq z_h$, $\alpha = 0$.

In the intermediate and low F_0 cases, the inefficient risk taking begins at z_{RN}, while in the high F_0 case it begins at z^{**}. The next two lemmas show how these risk-taking ranges vary with F_0.

LEMMA 5. *In the high F_0 case, the risk-taking range is shrinking as F_0 decreases.*

Proof. By Lemma 3, $\partial z^{**}/\partial F_0 > 0$. □

LEMMA 6. *In the intermediate and low* F_0 *cases, the risk-taking range is increasing as F_0 decreases.*

Proof. By Lemma 4, z_{rn} is rising as F_0 decreases. □

As F_0 decreases, the risk-taking range decreases in the intermediate case, but increases in the high and low cases. It is immediate that the optimal F_0 is on the boundary between the high and intermediate cases:

PROPOSITION 2. *The constrained socially optimal F_0 is such that* $z^{**} = z_{\text{RN}}$.

Figure 12.5 depicts the optimal configuration. The proposition results from the fact that any reduction in asset substitution brought about through renegotiation is welfare improving. Since the bank forgives over the range $[z^{**}, z^*]$, that range of borrowers is discouraged from inefficiently adding risk. Any higher value of F_0 would make it more costly on the margin for the bank to forgive sufficiently to prevent asset substitution. This would have the effect of raising z^{**} and thereby increasing the range of asset substitution. Any lower value of F_0 would increase z_{RN} that is, it would provide the bank with a credible threat to liquidate for the marginal borrower. The effect would be a transfer to the bank at the cost

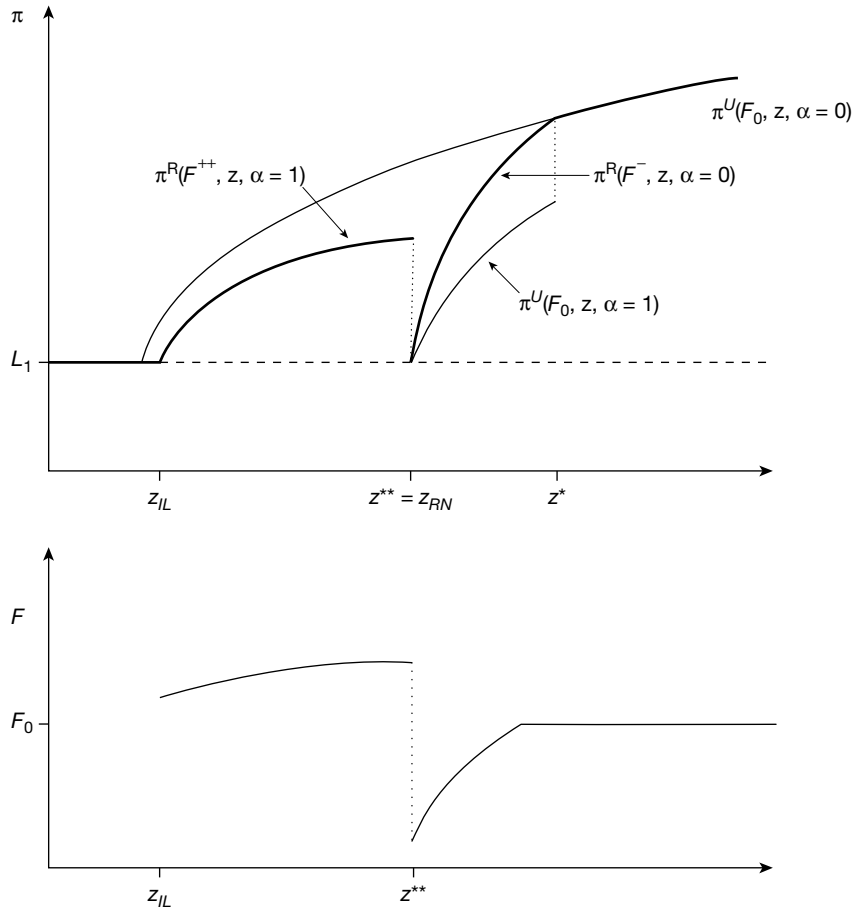


Figure 12.5

of a decline in project quality, as the bank's ability to behave opportunistically entices it to abandon its antiasset substitution measures. Thus the equilibrium F_0 optimally balances off the two moral hazard problems.

While we have yet to discuss how the socially optimal F_0 of Proposition 2 will be implemented, we stress the importance of the proposition. The face value of the debt serves a critical role in allocating bargaining power between borrowers and lenders. It would only be a complete coincidence if that face value bore any relation to default risk. Consequently there is no reason to expect the equilibrium F_0 to imply zero profits. The next section addresses this last issue.

12.4.2. Implementation of the Socially Optimal F_0

Let F_0^* denote the optimal value of F_0. Given the nature of bank loans, it should be clear that linear pricing is not necessary. Thus if F_0^* implied that banks would

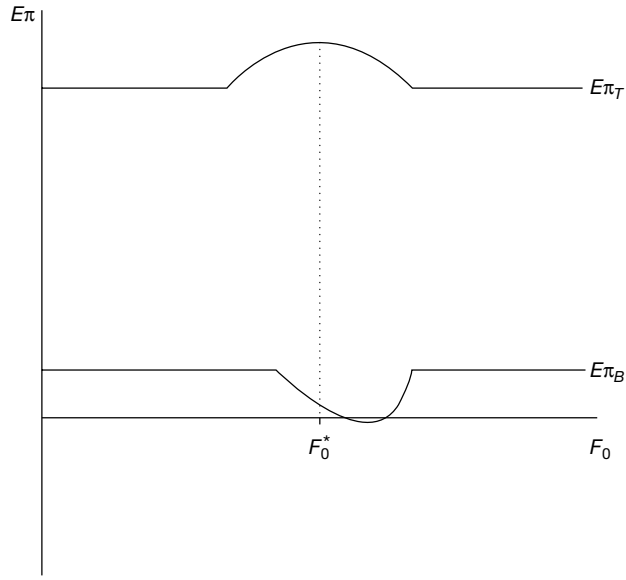


Figure 12.6

make positive profits, competitive banks could still price loans at F_0^* and compete by offering other goods or services for free, up to the point that they make zero profits on the whole package. This is the case depicted in figure 12.6. While it might seem odd that expected profits for the bank are declining at $F_0 = F_0^*$ the intuition should be clear: a lower value of F_0 would be regarded with suspicion by borrowers, who would foresee that the bank would be more likely to hold them up in the interim. Borrowers would thus prefer the slightly higher F_0 because it is more credible. Also, note that bank profits are the same at extreme values of F_0 because the range of risk taking is broad and the renegotiated F would be the same in either case. The point is that a very high or very low value of F_0 is ignored, as both sides know it will be reset in the interim.

On the other hand, it is possible that at $F_0 = F_0^*$ banks would make negative expected profits. In this case, competitive banks could charge origination fees to make up the difference, if that were feasible. Under our assumptions, however, the borrower has no surplus liquidity at $t = 0$, so competitive banking cannot implement the social optimum. In this case, the bank would have to lend the borrower additional money to cover the origination fee. But this would be tantamount to charging a higher F_0. Thus if F_0^* did imply negative profits, and there were no way to extract origination fees from the borrower without effectively increasing the borrower's leverage, then competition would drive F_0 to the zero-profit point (as the figure makes clear, there would likely be more than one) that had the highest total profits. This would be inefficient relative to the scenario depicted in figure 12.6, but would be the best the system could accomplish.

Our result that bank loans will generally involve nonlinear pricing is consistent with the observation that the loan rate is only one component of pricing bank loans. In addition to the interest rate, banks also use a variety of fees and, at least in the past, tied lending to other services. Booth and Chua (1995) discuss the prevalence of, and different types of fees in, bank loan contracts. For example, Booth and Chua find that an up-front fee is charged in 45% of the sample loan contracts examined. Other fees are not mutually exclusive and are also common. Overall, Booth and Chua show that substantial heterogeneity exists in the pricing of loan contracts. Our explanation for the presence of such pricing structures differs considerably from the existing literature. To explain this structure of bank loan pricing, the previous literature has focused on the presence of informational asymmetries related to the credit risk of the borrower. In Thakor and Udell (1987) borrowers reveal their default characteristics based on their choice of contract terms. In Berlin (1987), borrowers self-select across contract types based on their probability of borrowing.

12.5. DISCUSSION

What makes bank loans valuable? Why are bank loans senior? In this section we discuss how our model addresses these questions.

12.5.1. Bank Loans, the Option to Liquidate, and Corporate Bonds

The features of bank loans that distinguish them from conventional corporate bonds are the bank's ability to renegotiate the terms of existing loans and to call in or "liquidate" them if that is desirable. Thus, as mentioned above, there are really three distinct securities to consider: corporate bonds, bank loans with the liquidation option, and bank loans without the liquidation option. It should be immediately clear that the bank loan without the liquidation option dominates corporate bonds. Banks by assumption have the ability to renegotiate, which leads to more efficient outcomes in some states of the world. Otherwise there is no difference, so the gain in efficiency is unambiguous. Literally interpreted, this result would turn the question of bank loan's value on its head and raise the question of why corporate bonds are valuable. This result does not, however, immediately extend to junior corporate bonds issued in addition to bank loans (see discussion below). Moreover, in practice firms have begun to have corporate bonds mimic the forgiveness feature of bank loans by utilizing exchange offers. Typically such exchange offers are a device for the borrowing firm to initiate forgiveness and reduce its debt obligations [see Asquith, Mullins, and Wolff (1989)]. This is discussed by Gertner and Scharfstein (1991).

So the remaining question concerns the value of the liquidation option. Although we regard the liquidation option as virtually intrinsic to bank loans, it is still useful to analyze why such an option would be valuable. The value of this option hinges entirely on the range of projects in which liquidation occurs, that is, $[z_1, z_{\mathrm{IL}}]$. We know that in general there is (at least in the absence of side payments) inefficient liquidation over the range $[z_{\mathrm{EL2}}, z_{\mathrm{IL}}]$. On the other hand, in the absence of the liquidation option there would be inefficient *continuation* over the range $[z_1, z_{\mathrm{EL2}}]$.

Clearly, the desirability of the liquidation option is in general ambiguous. It will depend on the shape of the density of z over these ranges as well as the mapping from z to expected payoffs. The fact that bank loans almost invariably do contain a liquidation option (usually implicitly through covenants) suggests that the excessive liquidation costs may in practice be relatively small—perhaps because of side payments, but also because the range or magnitude of inefficient liquidations is simply not very large in comparison to the problem of excessive continuation in the absence of banks' ability to liquidate.

12.5.2. Junior Debt and Related Concerns

Our model does not explicitly include junior debt. There is little loss in generality though, because everything in the article carries through conditional on the presence of a fixed amount of junior debt associated with the project. Since bank loan covenants would generally specify limits on junior debt, the bank can simply consider junior debt as part of the borrower's project, and whatever agency problems may be associated with junior debt can be thought of as already accounted for in the probability distribution over project payoffs.

Even though we do not treat junior debt explicitly, our model nevertheless sheds light on a puzzle that emerges from the existing literature on financial intermediation: Why should banks as senior claimants engage in monitoring the behavior of borrowers more closely than junior claimants do? Junior claimants would seem to have a greater incentive to monitor (in a costly state verification setting), as Fama (1985) has argued.[16] Our view is that in addition to their ability to act unilaterally, banks' status as senior claimants puts them in the position to gain the most in the event of liquidation. Certainly junior creditors can

16. Fama (1985) argues that the benefits of banks' monitoring activities spill over into the corporate debt market as the presence of bank debt on a corporation's balance sheet functions as a sort of "seal of approval" that enables it to issue debt directly. The problem with this scenario is that bank debt is senior to corporate debt. Consequently banks should have less incentive to monitor borrowers' subsequent behavior than the junior creditors would have. Yet firms often have both bank loans and publicly issued and traded bonds.

force a borrower into bankruptcy, but then they risk getting little or nothing because of their junior status. Banks, as senior claimants, have an incentive to force liquidation, possibly excessively so, as we have seen. If the likelihood of excessive liquidation can be reduced via prepayment options, then bank loans dominate other forms of debt because the prospect of relatively efficient liquidation raises the value of the firm ex ante by lowering the cost of debt.[17,18] If senior creditors were decentralized they would find it costly to undertake the efficiency-enhancing renegotiation process to avoid asset substitution and inefficient liquidation.

The presence of decentralized junior debt could make it more difficult for the bank to preclude asset substitution through renegotiation, but there are ways around that. The difficulty is that the temptation to take on risk is a function of the debt:equity ratio. The bank would have to forgive more debt in order to counter the borrower's incentive to add risk if there are junior debtholders, and some of the benefits would spill over to them. Moreover, even if it is in the collective interest of the junior debtholders to participate in the forgiving, there is a free-rider problem, as each debtholder would try to hold out and let the others bear the burden.

One mechanism a bank has at its disposal to deal with the free-rider problem works as follows. The bank can say to the firm: "We will forgive $x\%$ of the debt provided you can get the junior debtholders to do so as well." The firm can appeal to the junior debtholders through a consent solicitation that amounts to a "coercive exchange offer" [see Kahan and Tuckman (1993)], which effectively plays off the junior debtholders against each other to get them to do what is in their collective interest. Kahan and Tuckman find that even though such consent solicitations involve apparent redistributions of wealth from bondholders to stockholders, they are typically associated with positive abnormal bondholder returns. This is consistent with the spirit of our analysis which argues that such renegotiations are efficiency enhancing. Of course the ability of firms to induce renegotiation with decentralized junior debtholders suggests that such renegotiation is not impossible, as we have assumed, but merely more costly than with banks.

17. Prepayment is another contract feature that we did not consider, but that works in favor of bank loans. A prepayment option allowing the borrower to prepay debt at date $t = 1$ can reduce the cost of excessive liquidation by the bank, increasing the benefits of loans over bonds. Then, as shown in Gorton and Kahn (1994), inefficient liquidation can be reduced or eliminated and borrowers might never want to add risk.

18. As junior claimants banks could still forgive debt, while as senior claimants they would not forgive since subordinated debtors would be the beneficiaries. Thus when junior debt is present, and banks are senior lenders, banks are not likely to forgive principal. This corresponds to the findings of Asquith, Gertner, and Scharfstein (1991) who study distressed junk bond issuers and find that the banks rarely forgave principal, but did defer principal and interest payments.

12.6. FINAL REMARKS

We summarize our key findings as follows:

1. Since the key advantage of bank loans arises from banks' ability to monitor and renegotiate in order to mitigate moral hazard problems, it is not surprising that the key determinant of bank loan pricing is also the mitigation of moral hazard. Specifically, we find that the equilibrium interest rate on loans does not primarily reflect a default premium. Rather, it is the rate that results ex ante in minimal expected asset substitution by borrowers. Since there is no guarantee that this rate results in zero profits, competition by banks will result in nonlinear (in the amount borrowed) pricing arrangements for loans.
2. The volatility of corporate securities is endogenous and variable. The firm sometimes has an incentive to increase volatility. The outside claimant that is in a position to prevent this, the bank, only imperfectly controls borrower risk-taking. The bank interacts with the borrower during the course of the contract. It is in a position to do this because by assumption it is a single agent and so can renegotiate higher interest rates, liquidate, or forgive debt. The bank controls risk in two ways: it may liquidate the project or it may change the borrower's incentive to add risk by debt forgiveness. But, importantly, there are borrower types for which the bank cannot prevent risk from being added, but whose projects are allowed to continue. This means that the variance of the value of the firm (and the mean) depend, in equilibrium, on the borrower type and, in particular, is not constant.
3. The social value of bank loans relative to other instruments presumes that excessive liquidation costs are small relative to excessive continuation costs, that is, that banks do not, in effect, "throw the baby out with the bath water" in the course of monitoring and liquidating projects.

APPENDIX A

Parameter Restrictions

The following assumptions involve an endogenous variable, F, and therefore must be handled with care. Their role is only to ensure that the parameters of the problem are such that the model behaves reasonably. It turns out that for extreme values of F the characterizations of outcomes in the paper are not complete. These additional cases are either implausible or economically uninteresting, and would only burden the article with additional complexity. The

essence of the assumptions is to show that these outcomes can be ruled out by appropriate (and mutually compatible) parameter restrictions.

Let F_0 denote the amount initially specified by the contract to be repaid at $t = 0$. Clearly F_0 must be in the range $[D, V_h, + \in_h,]$. At $t = 1$ a different amount, F^N, may be negotiated. Let F denote either of these values. Then:

Assumption 12. $\in_h \ > c + F$.

In other words, the upper bound of the support of $\in$ is sufficiently large that adding risk always results in a positive probability of solvency. This assumption simply makes the problem interesting since it says that when risk is added there is always some chance for the borrower to benefit.

Assumption 13. $L_2 + c/[1–K(c)] > F$.

(Recall that $K(\in)$ is the distribution function for $\in$.) This assumption says that c is sufficiently large and/or the distribution of $\in$ is sufficiently skewed that for a given F, the bank always prefers that the borrower not add risk. Again, this is the interesting problem since otherwise the bank would not want to prevent asset substitution.

Let $F^{\#}(\alpha, z) = \arg\max_F (L_2 - \alpha c)G(F|V; z, \alpha) + F(1–G(F|V; z, \alpha))$. This is the value of F that maximizes the bank's expected profit as of $t = 1$ for a borrower of type z. Let $F^{\#} = \inf\{F^{\#}(\alpha, z)\}$. Then:

Assumption 14. *$F^{\#}$ is larger than any F_0 or F^N that the bank would consider.*

This assumption ensures that bank profits are increasing in F over the relevant range. It is straightforward to extend the results of this article to the case where F_0 or F^N is larger than $F^{\#}$. Lenders can always forgive debt at $t = 1$ in order to ensure that they are on the upward sloping portion of the bank profit function. The assumption allows us to ignore this issue of forgiveness (which has no efficiency considerations). To avoid burdening the article with additional complexity, in what follows we will always assume that any F under consideration is less than $F^{\#}$.

Assumptions 12, 13, and 14 ensure that, whatever the equilibrium F turns out to be, we can choose parameters that are consistent with the characterizations in the analysis.

APPENDIX B

PROOF OF LEMMA 1. The first part of the lemma says that there exists a trigger value of z which we denote z^*, such that the borrower chooses $\alpha = 1$ if and only if $z \leq z^*$. That is, the moral hazard problem is more severe for those who get bad news. In the following discussion we use the notation $E_x[\omega(x,y)]$, where ω

is a function of random variables x and y, to indicate that the expectation is with respect to x alone. We first provide the following lemma.

LEMMA A1. *Let V and z be two random variables with joint distribution $G(V,z)$ and assume that the conditional distribution of z given V has the MLRP property. Let $\psi : R \to R$ be some continuous function that crosses zero only once, and from above. Then the function $\xi : R \to R \xi(z) = Ev[\psi(V, F)|z]$ crosses zero at most once, and from above.*

Proof. See Karlin (1968).

Recall that $\psi(V,F_0) \equiv E_{\in}\left[\pi^F(V+\in -c,F_0) - \pi^F(V,F_0)\right]$, where $\pi^F(\omega) = \max[\omega - F_0, 0]$ is the profit to the borrowing firm. We denoted the expected gain to a borrower of type z from switching from project $\alpha = 0$ to $\alpha = 1$ by $\Gamma(z)$. Hence $\Gamma(z) = Ev[\psi(V, F_0)|z]$. At $t = 1$, having observed z, the borrower chooses α to maximize profits. To prove Lemma 2 we apply Lemma A1 and need only show that $\psi(V,F_0)$ crosses zero only once, and from above. By Assumption 15, the upper bound of the support of $\in$ is greater than $c + F_0$. We have

$$\psi(V,F_0) = \int_{c+F_0-V}^{\in_h} [\in -(c+F_0-V)]\, h(\in)\, d\in = -\max[V-F_0, 0].$$

We know that $V \leq F_0$ implies $\psi(V,F_0) > 0$. Further, since for $V > F_0$

$$\psi(V, F_0) = \int_{c+F_0-V}^{\in_h} \in h(\in) d\in \\ -(c+F_0-V)(1-H(c+F_0-V)) - (V-F_0),$$

we have

$$\lim_{v\to\infty} \psi(V, F_0) = \lim_{v\to\infty} -VH(c+F_0-V) - c < 0.$$

We also have, for $V > Fo$,

$$\frac{\partial \psi}{\partial V} = -H(c+F_0-V) \leq 0.$$

Therefore, ψ has the desired properties, and we have proven the proposition. □

We now turn to proving the second part of the lemma, that is, that z^* is increasing in F_0. We have $\Gamma(z^*, F_0) = 0$ implicitly defining $z^*(F_0)$. To prove that z^* is increasing in F_0, it suffices to show that

$$-\frac{\partial \Gamma}{\partial F_0} \Big/ \frac{\partial \Gamma}{\partial z} > 0$$

evaluated at z^* and F_0. By the proof of Lemma 1, we already know that $\partial\Gamma(z^*)/\partial z < 0$, since at z^* the function Γ crosses zero from above. So it

remains to show that $\partial\Gamma(z^*, F_0)/F_0 > 0$. For this we need to see how $\psi(V, F_0)$ depends on F_0. We have from before,

$$\psi(V, F_0) = \int_{c+F_0-V}^{\in_h} \left[\in -(c + F_0 - V)h(\in)d\in - \max[V] - F_0, 0\right],$$

which we now want to consider as a function of F_0 holding V fixed. But it is straightforward to verify that $\partial\psi/\partial F_0 > 0$. Hence $\partial\Gamma(z^*, F_0)/\partial F_0 = E[\partial\psi(V, F_0)/\partial F_0 \,|z^*] > 0$.

PROOF OF LEMMA 2. Define the gain to the bank from the borrower of type z adding risk to be $\Gamma_B(z; F) = E_v[\omega(V)|z]$, where

$$\begin{aligned} \omega(V, F) &= -c + [1 - H(F + c - V)](F - L_2) && \text{if} \quad V < F \\ &= -H(F + c - V)(F - L_2 + c) && \text{if} \quad V \geq F. \end{aligned}$$

$\omega(V)$ is discontinuous at $V = F$. Also $\omega(V)$ can be positive for $V < F$ in the vicinity of F. But, for given F, $\omega(V) < 0$, for all V, if $F < L_2 + c/[1 - H(c)]$. This cannot be true for all possible values of F, but for any given value it suffices that c or $H(c)$ be sufficiently large. But Assumption 15 states that $\in_h > c + F$, and Assumption 16 states that $L_2 + c/[1 - H(c)] > F$. Thus $\omega(V)$ is assured of lying everywhere below zero. Recalling that ψ is the gain to the borrower, we have shown that $\psi + \omega$, which is the social gain, lies everywhere below ψ. □

PROOF OF PROPOSITION 1. We take the cases in reverse order. Part 4: First, we must show that $[z_{IL}, z_{RN}]$ exists. For $z > z_{IL}$, $\Gamma(z) > 0$ implies $\Pr(V > F_0) > 0$, that is, $\pi^T(z, \alpha = 1) > L_1$. That implies $\pi^U(F_0, z, \alpha = 1) > 0$. As $z \to z_{IL}$, $\pi^T(z, \alpha = 1) \to L_1$ and $\pi^U(F_0, z, \alpha = 1) < L_1$. Thus $[z_{IL}, z_{RN}]$ exists. In the interval $[z_{IL}, z_{RN}]$, $\pi^T(F^N, z, \alpha = 1) > L_1$, so the project should not be liquidated, but $\pi^U(F_0, z_{RN}, \alpha = 1) < L_1$, that is, at the unrenegotiated contract the bank would be better off liquidating the project. Thus, $F^N = F_0$ is not optimal. The fact that $z_{RN} < z^{**}$ means that $\pi^R(F^-, z, \alpha = 0) < \pi^U(F_0, z, \alpha = 1)$. Therefore, forgiving some of the debt by lowering the interest rate cannot be optimal. Hence, the project is profitable even if the borrower chooses $\alpha = 1$, and the bank sets $F^N = F^{++}(z)$, that is, raises the interest rate. Part 3: The borrower will choose $\alpha = 1$ because $z < z^*$, but the bank cannot raise the interest rate because it has no credible threat since $z > z_{RN}$. $\pi^R(F^-, \alpha = 0, z) < \pi^U(F_0, \alpha = 1, z)$ because $z < z^{**}$, so debt forgiveness is not optimal. Since $\pi^U(F_0, \alpha = 1, z) > L_1$, the best the bank can do is maintain the current contract. Part 2: In this range borrowers choose to add risk, $\alpha = 1$, since $z < z*$, but the bank has no credible liquidation threat since $z_{RN} < z^{**}$. However, assuming the interval $[z^{**}, z^*]$ exists, lowering the interest rate results in $\pi^R(F^-, z, \alpha = 0) > \pi^U(F_0, z, \alpha = 1)$. Part 1: Borrowers in this range do not add risk and the bank has no credible threat. Thus the best the bank can do is maintain the initial contract. □

APPENDIX C

Renegotiation outcomes for the intermediate F_0 case

The intermediate F_0 case is the situation where $z^{**} < z_{RN} < z^*$. Liquidation occurs for $z < z_{IL}$.

PROPOSITION B1. *If $z^{**} < z_{RN} < z^*$. then renegotiation results in:*

(i) $F^N(z) = F^+(z) > F_0$, *for all* $z \in [z_{IL}, z_{RN}]$; *that is, raise rate; borrower chooses* $\alpha = 1$.

(ii) $F^N(z) = F^-(z) < F_0$, *for all* $z > z_{RN}$; *that is, forgive debt; borrower chooses* $\alpha = 0$.

Proof. Part 1: For $z \in [z_{IL}, z_{RN}]$ the borrower will choose $\alpha = 1$, ceteris paribus. Liquidation is not optimal for these borrowers since $z > z_{IL}$. Also, because $z < z_{RN}$, $\pi^U(F, z, \alpha = 1) < L_1$, so maintenance of the initial contract is not optimal. Since $z < z_{RN}$ the bank can credibly threaten the borrower. By Assumption 11, $\pi^R(F^{++}, z, \alpha = 1) > \pi^R(F^+, z, \alpha = 0)$, that is, it is more profitable for the bank to raise the rate by so much that the borrower chooses $\alpha = 1$, rather than raise the rate to the point where the maximum surplus is extracted and the borrower chooses $\alpha = 0$. So the bank raises the interest rate and the borrower chooses $\alpha = 1$. Part 2: For $z > z^*$ the project is profitable and the borrower will choose $\alpha = 0$, ceteris paribus. The bank cannot threaten the borrower since $z_{RN} < z^*$, so the initial contract is maintained. □

Renegotiation Outcomes for the Low F_0 Case

The low case is the situation where $z^{**} < z^* < z_{RN}$, that is, unrenegotiated bank profits are less than the liquidation value starting at borrower types higher than the type at which there is an incentive to switch projects and add risk. In this situation the bank can credibly threaten to liquidate borrowers who have no intention of switching projects (in addition to those who do).

PROPOSITION B2. If $z^{**} < z^* < z_{RN}$, then renegotiation results in the following outcomes:

(i) $F^N(z) = F^+(z) > F_0$ *for all* $z \in [z_{IL}, z_{RN}]$; *that is, raise rate; borrower chooses* $\alpha = 1$;

(ii) $F^N(z) = F_0$ *for all* $z > z_{RN}$; *that is, no change; borrower chooses* $\alpha = 0$.

Proof. Similar to Proposition Bl. □

Alternatives to Assumption 11

Assumption 11 assumed that $\pi^R(F^N, z, 1) > \pi^R(F^N, z, 0)$ for all z and F. We now briefly reconsider Propositions 1, Bl, and B2, when Assumption 11 is

not assumed. The first alternative to Assumption 11, subcase 1, occurs when $\pi^R(F^N, z, 1)$ cuts $\pi^R(F^N, z, 0)$ from above at a point $\hat{z}$ such that $z_{\text{IL}} < \hat{z} < z_{\text{RN}}$. For this case:

PROPOSITION B3. *If $z_{\text{RN}} < z^{**} < z^*$, and subcase 1, then renegotiation results in:*

(i) $F^N(z) = F^{++}(z) > F_0$, *for all* $z \in [z_{\text{IL}}, \hat{z}]$; *that is, raise rate; borrower chooses* $\alpha = 1$.

(ii) $F^N(z) = F^{+}(z) < F_0$, *for all* $z \in [\hat{z}, z_{\text{RN}}]$; *that is, raise the rate but such that the borrower chooses* $\alpha = 0$.

(iii) $F^N(z) = F_0$, *for all* $z \in [z_{\text{RN}}, z^{**}]$; *that is, no change; borrower chooses* $\alpha = 1$.

(iv) $F^N(z) = F^{-}(z)$, *for all* $z \in [z^{**}, z^*]$; *that is, forgive debt; borrower chooses* $\alpha = 0$.

(v) $F^N(z) = F_0$, for *all* $z > z^*$; that is, no *change; borrower chooses* $\alpha = 0$.

Proof. Part 1: For $z \in [z_{\text{IL}}, \hat{z}]$ the borrower is choosing $\alpha = 1$. Liquidation is not optimal since $z > z_{\text{IL}}$. Since $\hat{z} < z_{\text{RN}}$, $\pi^U(F, z, \alpha = 1) < L_1$, so maintenance of the initial contract is not optimal. By the definition of subcase 1, $\pi^R(F^{++}, z, \alpha = 1) > \pi^R(F^{+}, z, \alpha = 0)$ so the bank raises the interest rate. Part 2: As above, neither liquidation nor maintenance of the initial contract is optimal. But, in this range, by the definition of subcase 1, $\pi^R(F^{++}, z, \alpha = 1) < \pi^R(F^{+}, z, \alpha = 0)$ so the bank raises the rate as far as possible while maintaining the incentive for the borrower to choose $\alpha = 1$. Part 3: In this range the bank can no longer credibly threaten the borrower so raising the rate is not feasible. Forgiveness is not profitable for the bank (by definition of z^{**}). So the rate does not change and the borrower chooses $\alpha = 1$. Part 4: Now it is profitable to forgive debt so that the borrower chooses $\alpha = 0$. Part 5: In this range the borrower will choose $\alpha = 0$, ceteris paribus. The bank has no credible threat to liquidate and cannot raise the rate. The rate stays the same and the borrower chooses $\alpha = 0$. □

Subcase 2 is the situation where $z_{\text{RN}} < \hat{z} < z^{**} < z^*$. In this case, the result is the same as above since the bank cannot threaten to liquidate borrowers of type $\hat{z} \in [\hat{z}, z_{\text{RN}}]$. Subcase 3 is $z_{\text{RN}} < z^{**} < \hat{z} < z^*$. Again, there is no change, for the same reason. The same is true for the case where $z_{\text{RN}} < z^{**} < z^* < \hat{z}$. The final possibility is the case where $\pi^R(F^N, z, 1) < \pi^R(F^N, z, 0)$ for all z and F, the opposite assumption of Assumption 11. In this case, it can easily be shown that the borrower never adds risk, since it is always profitable for the bank to forgive rather than raise the rate.

For the intermediate and low F_0 cases there are similar, straightforward variations when we deviate from Assumption 11. These are omitted for the sake of space.

REFERENCES

Asquith, P., R. Gertner, and D. Scharfstein, 1991, "Anatomy of Financial Distress: An Examination of Junk Bond Issuers," *Quarterly Journal of Economics,* 109, 625–58.

Asquith, P., D. Mullins, and E. Wolff, 1989, "Original Issue High Yield Bonds: Aging Analyses of Defaults, Exchanges, and Calls," *Journal of Finance,* 44, 923–52.

Berlin, M., 1987, "Loan Commitments: Insurance Contracts in a Risky World," Federal Reserve Bank of Philadelphia, *Business Review,* 3–12.

Bolton, P., and D. Scharfstein, 1990, "A Theory of Predation Based on Agency Problems in Financial Contracting," *American Economic Review,* 54, 525–40.

Booth, J., and L. Chua, 1995, "Structure and Pricing of Large Bank Loans," Federal Reserve Bank of San Francisco, *Economic Review,* 3, 52–62.

Detragiache, E., 1994, "Public versus Private Borrowing: A Theory with Implications for Bankruptcy Reform," *Journal of Financial Intermediation,* 3, 327–54.

Diamond, D., 1984, "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies,* LI, 393–414.

Fama, E., 1985, "What's Different About Banks?," *Journal of Monetary Economics,* 15, 29–40.

Gertner, R., and D. Scharfstein, 1991, "A Theory of Workouts and the Effects of Reorganization Law," *Journal of Finance,* 46, 1189–1222.

Gilson, S., K. John, and L. Lang, 1990, "Troubled Debt Restructurings: An Empirical Study of Private Reorganization of Firms in Default," *Journal of Financial Economics,* 27, 315–54.

Gorton, G., and J. Kahn, 1992, "The Design of Bank Loan Contracts, Collateral, and Renegotiation," Working Paper 4273, NBER, Cambridge, MA.

Hart, O., and J. Moore, 1998, "Default and Renegotiation: A Dynamic Model of Debt," *Quarterly Journal of Economics,* 113, 1–41.

Hoshi, T., A. Kashyap, and D. Scharfstein, 1990, "The Role of Banks in Reducing the Costs of Financial Distress in Japan," *Journal of Financial Economics,* 27, 67–88.

James, C., 1987, "Some Evidence on the Uniqueness of Bank Loans," *Journal of Financial Economics,* 19, 217–36.

Kahan, M., and B. Tuckman, 1993, "Do Bondholders Lose from Junk Bond Covenant Changes?" *Journal of Business,* 66, 499–516.

Kahn, J., 1992, "Debt, Asymmetric Information, and Bankruptcy," working paper, University of Rochester.

Karlin, S., 1968, *Total Positivity,* Stanford University Press, Palo Alto, CA.

Lummer, S., and J. McConnell, 1989, "Further Evidence on the Bank Lending Process and the Capital Market Response to Bank Loan Agreements," *Journal of Financial Economics,* 25, 99–122.

Milgrom, P., 1981, "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics,* 12, 380–91.

Morsman, E., 1986, "Commercial Loan Structuring," *Journal of Commercial Bank Lending,* 68.

Quill, G., J. Cresci, and B. Shuter, 1977, "Some Considerations in Secured Lending," *Journal of Commercial Bank Lending,* 59, 41–56.

Rajan, R., 1992, "Insiders and Outsiders: The Choice Between Informed and Arm's-Length Debt," *Journal of Finance,* 47, 1367–1400.

Rajan, R., and A. Winton, 1995, "Covenants and Collateral as Incentives to Monitor," *Journal of Finance,* 50, 1113–1146.

Sharpe, S., 1990, "Asymmetric Information, Bank Lending, and Implicit Contracts: A Stylized Model of Customer Relationships," *Journal of Finance,* 45, 1069–1087.

Slovin, M., M. Sushka, and J. Polonchek, 1993, "The Value of Bank Durability: Borrowers as Bank Shareholders," *Journal of Finance,* 48, 247–66.

Thakor, A., and G. Udell, 1987, "An Economic Rationale for the Pricing Structure of Bank Loan Commitments," *Journal of Banking and Finance,* 11, 271–89.

Zimmerman, C., 1975, "An Approach to Writing Loan Agreement Covenants," *Journal of Commercial Bank Lending,* 58, 2–17.

13

Universal Banking and the Performance of German Firms*

GARY B. GORTON AND FRANK A. SCHMID ■

13.1. INTRODUCTION

German universal banks appear to be powerful institutions in that they can own blocks of equity and vote individual shareholders' votes in proxy. This system has been controversial for over a century (e.g., Hilferding, 1910) and is addressed more recently in the report of the Gessler Commission (e.g., Studienkommission, 1979; and Krümmel, 1980), but apart from Cable (1985) there has been no empirical analysis of this corporate governance system and there is certainly no agreement about the effects of German banks on the performance of firms.

One view of the German system is that German banks are large, active, informed investors that improve the performance of firms to the extent that they hold equity and have voting power from casting the votes of small investors in proxy. Banks are seen as long-term investors who oversee firms' investments and organize internal capital markets, rather than acting as myopic investors (e.g., Porter, 1992; Grundfest, 1990). The banking relationship mitigates the costs of both external financing and of actively monitoring management. Proponents of this view see German banks as a model of active block shareholders

* Thanks to Anup Agrawal, Jörg Borrmann, William Cleveland, Bill Emmons, Silverio Foresi, Javier Hidalgo, Chris James, Shmuel Kandel, Mark Lang, Erich Loitlsberger, Claus Niemann, Benedikt Pötscher, Ragu Rajan, Reinhard H. Schmidt, René Stulz, and Andrei Shleifer (the referee) for suggestions and discussions. Also, thanks to Lori Gorton, Tatjana Greil, Helge Hagge, Thomas Hansen, Joachim Pansgerau, Ruth Paschka, and Martina Venz for research assistance. Gorton thanks the Bank of England for support during his tenure as a Houblon-Norman Fellow. Schmid thanks Deutsche Forschungsgemeinschaft for support when visiting the Wharton Financial Institutions Center. The views expressed in this paper are those of the authors and not necessarily those of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

that should be emulated in stock-market-based economies (where shareholders are dispersed and institutional investors are passive). For example, Grundfest (1990) asserts: "In Germany, large banks and industrial combines exercise substantial influence over the operation of many companies and are able to effect management and strategic changes when circumstances warrant" (p. 105).

Critics of universal banking see the enormous power of banks as harmful because of conflicts of interest that a bank faces when it simultaneously is a large equity holder in the firm, is in control of a large number of proxy votes, controls access to external capital markets, and has loans outstanding to the firm. Because banks themselves seem impervious to external control, the concentration of power in banks is seen as allowing them to essentially run firms in their own interests. For example, banks can refuse to allow cash to be paid out of firms in order to maintain "hidden reserves." Or a bank might force a value-reducing merger between a distressed and a nondistressed firm, both of which it controls. Wenger and Kaserer (1998) express this unfavorable view on German banks:

> . . . German banks do not only provide industrial companies with loan capital but also exercise considerable voting power in stockholder meetings of many public corporations. This is partly due to proxies of their clients and partly due to stock ownership. . . . we would argue that this specific institutional environment does not reduce agency problems; on the contrary, this situation is prone to enlarge and perpetuate these problems (p. 50).

Banking laws in Germany do not legally restrict commercial banks from holding blocks of equity in nonfinancial firms. Consequently, banks can have control rights in the form of votes that they would not have in the U.S., for example. As we will see below, however, bank blockholding is not so pervasive in Germany, while blockholding by nonbanks is extensive. The control rights of these blockholders can be limited by voting restrictions. For example, the voting rights of shareholders can be restricted by the firm's charter to a maximum fraction in the firm's total voting stock, regardless of the fraction of shares owned. While voting restrictions apply to any shareholder, banks can potentially exercise more votes because voting restrictions generally do not apply to votes that banks cast on behalf of small shareholders. For example, a firm can be owned by a single bank with 5% of the shares, a nonbank blockholder with 50% of the shares, and dispersed shareholders with the remainder. If there is a voting restriction constraining the votes of the nonbank blockholder to 10%, and if the bank further controls all of the proxy votes of the small shareholders, then the bank, in the absence of any other considerations, effectively controls this firm. (Changes to the firm's charter typically require a 75% majority.) Note that this could occur even if the bank owned no shares. In such a case, there is no link between cash-flow rights and control rights.

It is not only the role of German banks that has been controversial. There is an extensive literature on codetermination, that is, the laws requiring that firm employees hold voting seats on the supervisory boards of large firms. (In Germany, limited liability companies have a two-tiered board system.) Because of codetermination, governance of German firms does not depend solely on possession of control rights in the form of votes attached to equity shares. The controversy emanates from the ideological implications of dictating that some of the owners' control rights effectively be ceded to labor. Codetermination, for example, means that a large firm owned by a single shareholder, or perhaps a family, cannot appoint all the directors on the supervisory board. Under the two-tiered board system, management is insulated, at least to some extent, from discipline by shareholders. While the literature on German codetermination is massive, there is relatively little quantitative work assessing the impact of codetermination on firm performance; Gorton and Schmid (1998) provide a brief survey.

The theoretical effects on firms of the codetermination system are difficult to assess because the objectives of the employees are not obvious. On one hand, to the extent that employees are residual claimants by virtue of their investment of, possibly, firm-specific human capital, they will govern in the interests of shareholders. On the other hand, if their human capital is not diversifiable, risk-averse employees' objectives can differ from those of shareholders. In essence, codetermination reduces the value of control rights from equity ownership. In fact, Gorton and Schmid (1998) find that with employees on a firm's board, firm resources are directed to less productive uses, decreasing the return on assets, the return on equity, and the market-to-book ratio of equity.

Universal banking, proxy voting, and codetermination suggest that, in reality, corporate governance in Germany is much different from the system described by received theory (see La Porta, Lopez-de-Silanes, and Shleifer, 1999). In theory, corporate governance is based on the system of one share, one vote, an apparently incentive-compatible way of linking claims on cash flows with control rights. (Grossman and Hart, 1988, and Harris and Raviv, 1988, provide the theoretical arguments for the optimality of one share, one vote.) Germany, however, is clearly different from that model. Little is known about the German system due to a lack of theory rich enough to provide predictions in such a complicated setting, as well as a lack of data. Disclosure requirements in Germany simply do not exist to the same extent as in Anglo-American stockmarket-based economies. Nevertheless, in this paper we empirically investigate corporate governance in Germany. We study four data sets covering 1975 and 1986, each with different advantages and disadvantages.

An empirical description of the effects of the above corporate governance characteristics on the performance of German firms requires that we distinguish between equity ownership per se and the control rights that are derived from it. We need measures of control rights and control rights concentration, which we

can link to firm performance by some functional relation. Each of these steps is fraught with difficulty. With respect to control, one measure of control or power is the number of votes controlled by ultimate shareholders, following La Porta, Lopez-de-Silanes, and Shleifer (1999). Measuring control rights concentration requires a theoretical model of how large shareholders interact. While such models exist, they are based on voting behavior that implicitly assumes that cash-flow rights and control rights are closely linked. Moreover, these models cannot accommodate blockholders with different information, proxy voting, and voting restrictions. As we discuss below, we adopt the Herfindahl index as a measure of concentration that can be applied to the German case. Firm performance is not straightforward to measure either. Since Germany is less reliant on the stock market and has fewer disclosure requirements, we face the choice of relying on (German) accounting measures of performance or on market-based measures. The latter choice requires us to restrict our attention to publicly traded firms, an assumption that seems counter to the spirit of the investigation. We therefore use both accounting-based and market-based measures of performance.

There is also little theoretical guidance about the functional link between equity ownership and firm performance once the connection between cash-flow rights and control rights has, at least to some extent, been broken. Even for the more straightforward case of one share, one vote, as in the U.S., the relation between firm performance and the ownership stake of management has been argued to be nonlinear. Morck, Shleifer, and Vishny (1988), for example, examine the effect of insider concentration (the fraction of firm equity owned by top management) on nonfinancial firms' performance measured by Tobin's Q and find a piecewise linear, U-shaped relation. See also McConnell and Servaes (1990), who also examine U.S. nonfinancial firms, and Gorton and Rosen (1995), who analyze U.S. banks.

The German case is even more complicated than the U.S. case. While it is clear that the more cash-flow rights in a firm a party has, the more this party will want to improve the firm's performance, it is not clear what the objective function is for a party with control rights substantially in excess of cash-flow rights. This party might be interested in extracting private benefits rather than improving the value of cash-flow rights to which it has only a small claim. Thus, an important difficulty with analyzing the effects of banks on firms in Germany is that the bank can face conflicts of interest over some ranges of bank equity holdings, proxy-voting, and other (i.e., nonbank) shareholdings, but not over other ranges. Moreover, voting restrictions clearly can have an impact. But aside from considerations of the distribution of effective voting power in relation to cash-flow rights, codetermination undermines the power of votes attached to equity shares. The power of banks, to the extent that it is not derived from ownership in voting stock, can further undermine equity control rights.

In our empirical investigation of the influence of German universal banks and codetermination on the performance of German firms, we take into account banks' control rights that emanate from ownership of voting stock, banks' proxy-voting rights, the concentration of control rights from equity ownership, and voting restrictions. Equity ownership can involve pyramids, cross-shareholdings, and stocks with multiple votes. Because of the complexity of the firm's control structure, we test semiparametric specifications against various parametric specifications to determine the appropriate shape of the relation. This allows us to test for conflicts of interest between firm shareholders and banks, and between employees and shareholders. Further structure is then imposed in the form of a parametric specification. We also examine the influence of banks and employees on boards of directors.

The paper proceeds as follows. In Section 13.2 we describe the samples and discuss issues concerning the measurement of control rights in Germany. We also discuss the construction of variables that will be used in econometric tests. In Section 13.3 we propose hypotheses. Section 13.4 outlines the econometric methodology. Section 13.5 presents the basic set of results. Section 13.6 analyzes banks' representation on corporate boards. Section 13.7 is a discussion of the results. Section 13.8 is a brief conclusion.

13.2. MEASURING CONTROL RIGHTS, CONTROL RIGHTS CONCENTRATION, AND THE PERFORMANCE OF GERMAN FIRMS

Four issues are critical to our empirical analysis. First, we must construct a measure of equity control rights from data on ownership of (voting) stock. Second, we need a measure of concentration of the equity control rights. Third, we need a measure of firm performance. Finally, we need a functional specification for the link between control rights, control rights concentration, and firm performance. In this section we introduce the data sets. We then discuss two of the three measurement issues. We summarize the equity control rights structure of German firms based on our samples and we discuss voting restrictions. Finally, we address the third measurement issue and discuss firm performance measures and some other variables that we will use later.

13.2.1. Data Samples

Our data sets, discussed in detail in Appendix A, consist of four cross-sections of large public limited companies known as Aktiengesellschaften (AGs). For each of the years 1975 and 1986 we have a small sample and a large sample. The

German economy has been changing rapidly in the last decade, and possibly earlier as well. In order to study the economy prior to these changes, we start as far back as data availability will reasonably allow, i.e., 1975, but then include samples from ten years following in order to see if there are changes over the period 1975–1986.

The small samples are restricted in size due to the costs of collecting data on proxy voting. Furthermore, not all of the firms in the small samples are publicly traded. The small samples consist of 82 firms in 1975 and 56 firms in 1986. When restricted to firms with traded equity, the sample sizes are 54 and 42, respectively. The large samples consist of 283 firms in 1975 and 280 in 1986, all publicly traded. The small samples enable us to study the effects of proxy voting; for the large samples, proxy voting information is not available.

13.2.2. Measuring Control Rights

It is not obvious how to measure control in Germany. The issue is complicated, first of all, because pyramiding, cross-shareholding (or circular ownership) and stocks with multiple votes separate cash-flow rights from control rights in the form of votes. Franks and Mayer (2000) and Emmons and Schmid (1998) discuss these structures in Germany while Wenger and Kaserer (1998) discuss the legal background. La Porta, Lopez-de-Silanes, and Shleifer (1999) argue that a measure of control or power should be based on control rights that emanate from voting shares. We proceed similarly and calculate the control rights held by different parties, as explained below. It is not clear, however, that this procedure accurately defines control because of other complications besides pyramiding, cross-shareholding, and the existence of stocks with multiple votes. For example, as mentioned above, equity ownership is not the only legal basis for control because, under the system of codetermination, employees have votes on the supervisory board that are unrelated to holding shares. Thus, our strategy is to follow the concept of La Porta, Lopez-de-Silanes, and Shleifer, while taking account of all the other dimensions of governance with additional variables.

Cross-shareholding occurs when firms hold shares in each other, either directly or indirectly. An example of indirect cross-shareholding would be a triangular ownership structure with Firm A owning a block of Firm B's equity, Firm B owning a block of Firm C, and Firm C holding a stake in Firm A. There is a notable network of (mainly indirect) cross-shareholdings centered on Allianz AG, Germany's largest insurer (Wenger and Kaserer, 1998). This network comprises predominantly financial services firms. Outside this network, there are rare cases of cross-shareholdings, mainly among government-controlled utilities. In our samples (which exclude financial services firms), cross-shareholdings are not significant, as shown below.

Pyramiding occurs when Firm A owns a stake in Firm B, which owns a stake in Firm C. La Porta, Lopez-de-Silanes, and Shleifer (1999) define a pyramid as a chain of firms in which the chain includes at least one publicly traded company between the sample firm and the ultimate owners. (We discuss the notion of an "ultimate owner" below.) This definition will not suffice for Germany, as the middle firms in pyramids are almost invariably not traded. The typical case of pyramiding in Germany is joint ownership of nonfinancial firms, banks, or insurers in a financial holding shell (called Vermögensverwaltungs-, Vorschalt- or Beteiligungsgesellschaften) that holds a (controlling) stake in the sample firm. An example is Mercedes-Automobil-Holding AG, which (before it was dissolved in 1994) held a controlling stake in Daimler-Benz AG and was owned by a multitier shareholder structure that consisted mainly of financial firms (Franks and Mayer, 2000). Typically, a financial holding shell is not traded, has few or zero employees, exists solely to hold the stock of another firm, and has two to four owners, among them banks and insurance companies. In the case of Germany we say that pyramiding occurs when the sample firm's stock is held indirectly via (one or more) financial holding shells.

Figure 13.1 shows a typical example of a pyramid in our samples. Following our principle of deriving control rights from votes, the graph displays ownership

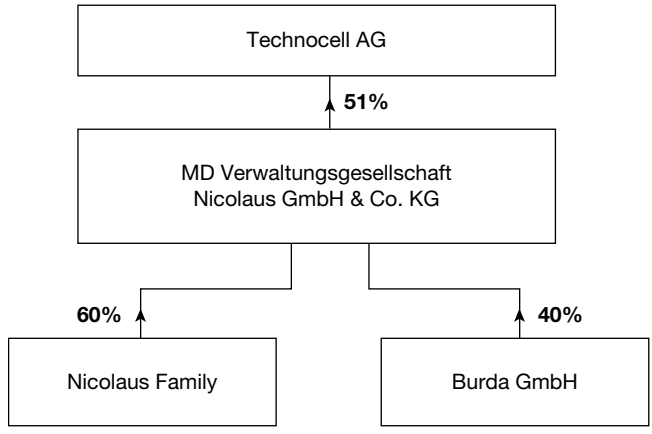


Figure 13.1 Technocell AG as an example of a simple pyramid, September 1986. Following our principle of defining control rights based on votes, the graph displays ownership as fractions of votes (which is not necessarily identical to the fractions of equity from which these votes emanate). Technocell AG has one blockholder, MD Verwaltungsgesellschaft Nicolaus GmbH & Co. KG, which owns 51%. This company, in turn, is owned by the Nicolaus family, with 60%, and by Burda GmbH, with 40%. MD Verwaltungsgesellschaft Nicolaus GmbH & Co. KG and Burda GmbH are not publicly traded. In this example, the Nicolaus family and Burda GmbH are the ultimate owners. The control rights are allocated as follows. The Nicolaus family holds 51% of Technocell and Burda GmbH holds 40%. That is, control rights are assigned based on the weakest link in the chain (La Porta et al., 1999a). Data source: *Saling Aktienführer* 1987, Verlag Hoppenstedt, Darmstadt 1986.

as percentages of votes (which is not necessarily identical to the percentages of equity these votes emanate from). Technocell AG has one blockholder, MD Verwaltungsgesellschaft Nicolaus GmbH & Co. KG, which owns 51%. This company, in turn, is owned by the Nicolaus family, with 60%, and by Burda GmbH, with 40%. MD Verwaltungsgesellschaft Nicolaus GmbH & Co. KG and Burda GmbH are not publicly traded. In this example, control rights are allocated as follows. The Nicolaus family holds 51% of Technocell and Burda GmbH holds 40%. That is, control rights are assigned based on the weakest link in the chain (La Porta, Lopez-de-Silanes, and Shleifer, 1999).

Few firms have stocks with multiple votes. While it has long been illegal in Germany to issue such stocks, those that existed prior to the change in legislation were grandfathered. There are only a few firms in our sample that have stocks with multiple votes, such as RWE AG and Siemens AG. In the case of RWE, a utility, provincial and municipal authorities hold stock that is endowed with 20 votes per share. In the case of Siemens, the family holds stock with six votes per share in certain decisions (as determined by the company charter). When we calculated control rights, we did so based on number of votes, not on number of stocks. In the case where multiple votes apply in certain circumstances only, such as with Siemens, we assumed the multiple-votes case.

Determination of control rights in complicated ownership structures (such as pyramids and circular ownership) depends on a definition of the *ultimate owner,* the agent at which tracing the ownership structure stops. We categorize firms into the following ultimate owners: banks (domestic and foreign), insurance companies (domestic and foreign), families and family trusts (domestic and foreign), government and government trusts (domestic and foreign), foreign financial holding shells (ownership data on these companies are generally not available), nonfinancial firms (domestic and foreign; no financial holding shells), and the sample firm itself (in the case of circular ownership). This classification of ultimate owners follows La Porta, Lopez-de-Silanes, and Shleifer (1999) except that we include nonfinancial firms as ultimate owners. This is because we often reach a point in the chain at which we cannot trace the holdings further because the (nonfinancial) firms are not publicly traded or there are insufficient data to determine the control rights structure. Recall that our samples are from the 1970s and 1980s, periods during which ownership data are sparse. Clearly, there is a certain arbitrariness to this procedure, but this is dictated by the data limitations that emanate from studying an economy that is not (at least during our sample periods) centered on the stock market. In the same vein, La Porta, Lopez-de-Silanes, and Shleifer do not break up firms that are not publicly traded, presumably because of a lack of data.

Table 13.1 shows the extent of pyramids, cross-shareholding, and circular shareholding in our samples. The table also shows the classification of our four samples into ultimate owners with a 25% cutoff rule. (The cutoff rule is based on control rights; it is applied for illustration and used in this table only; it is not

Table 13-1. Ultimate owners based on control rights that emanate from equity ownership. We follow La Porta et al. (1999a) when applying a cutoff rule to control rights (i.e., shareholders that control a smaller fraction of votes are not taken into account). Such a cutoff rule is employed in this table only, but not in subsequent tables or the empirical analysis. The cutoff rule applies to items 2, 3, and 4. We chose 25% as the cutoff level because this is an important threshold in Germany, as changes to the firm's charter generally require a 75% majority. The types of ultimate owners (item 3) and the types of largest ultimate owners (item 4) are not mutually exclusive because of the possibility of ties. Panel A: small samples. Panel B: large samples.

Equity Ownership Types	**1975 Sample**	**1986 Sample**
Panel A		
(1) Multi-Level equity ownership		
Pyramids (total)	10 (12%)	11 (20%)
Same shareholder owns directly and through pyramid	2 (2%)	2 (4%)
Circular ownership	1 (1%)	0 (0%)
(2) Existence of ultimate owners		
Ultimate owner exists	51 (62%)	39 (66%)
No ultimate owner exists	31 (38%)	19 (34%)
(3) Types of ultimate owners		
Banks (domestic or foreign)	24 (29%)	22 (39%)
Insurers (domestic or foreign)	7 (9%)	11 (20%)
Family, incl. family trusts (domestic or foreign)	11 (13%)	9 (16%)
Government (domestic or foreign)	13 (16%)	10 (18%)
Foreign financial holding shells	0 (0%)	2 (4%)
Nonfinancial firms (domestic or foreign)	25 (30%)	22 (39%)
Firm itself (circular ownership)	1 (1%)	0 (0%)
(4) Types of largest ultimate owner		
Banks (domestic or foreign)	20 (24%)	9 (16%)
Insurers (domestic or foreign)	7 (9%)	4 (7%)
Family, incl. family trusts (domestic or foreign)	11 (13%)	7 (13%)
Government (domestic or foreign)	12 (15%)	10 (18%)
Foreign financial holding shells	0 (0%)	2 (4%)
Nonfinancial firms (domestic or foreign)	19 (23%)	11 (20%)
Firm itself (circular ownership)	0 (0%)	0 (0%)
(5) Sample size		
Total number of firms	82	56
Number of publicly traded firms	54	42
Panel B		
(1) Multi-level equity ownership		
Pyramids (total)	12 (4%)	22 (8%)
Same shareholder owns directly and through pyramid	2(1%)	4 (1%)
Circular ownership	1 (0%)	1 (0%)
(2) Existence of ultimate owners		
Ultimate owner exists	238 (84%)	226 (81%)
No ultimate owner exists	45 (16%)	54 (19%)

Table 13-1. (CONTINUED)

Equity Ownership Types	**1975 Sample**	**1986 Sample**
(3) Types of ultimate owners		
Banks (domestic or foreign)	83 (29%)	61 (22%)
Insurers (domestic or foreign)	18 (6%)	18 (6%)
Family, incl. family trusts (domestic or foreign)	56 (20%)	77 (28%)
Government (domestic or foreign)	17 (6%)	21 (8%)
Foreign financial holding shells	1 (0%)	6 (2%)
Nonfinancial firms (domestic or foreign)	161 (57%)	147 (53%)
Firm itself (circular ownership)	1 (0%)	0 (0%)
(4) Types of largest ultimate owner		
Banks (domestic or foreign)	65 (23%)	34 (12%)
Insurers (domestic or foreign)	14 (5%)	8 (3%)
Family, incl. family trusts (domestic or foreign)	49 (17%)	67 (24%)
Government (domestic or foreign)	15 (5%)	18 (6%)
Foreign financial holding shells	0 (0%)	5 (2%)
Nonfinancial firms (domestic or foreign)	135 (48%)	117 (42%)
Firm itself (circular ownership)	0 (0%)	0 (0%)
(5) Sample size		
Total number of firms	283	280
Number of publicly traded firms	283	280

used in the subsequent quantitative analysis.) La Porta, Lopez-de-Silanes, and Shleifer (1999) introduce such a cutoff rule to isolate the shareholders in control from those not in control. We define the cutoff level to be 25% because corporate charters in Germany make this percentage a powerful block.[1] The ultimate owner with the largest fraction of control rights is deemed the largest ultimate owner, but there can be more than one such "largest" ultimate owner because of ties. With respect to types of ultimate owners, there are no appreciable differences in the samples between the two years analyzed. In our large samples, less than 20% of the firms are widely held, even less than in La Porta, Lopez-de-Silanes, and Shleifer (1999), who use a 20% cutoff rule and find that 50% of the firms in Germany are widely held. In our small samples, roughly 35% of the firms are widely held.

13.2.3. Measuring Concentration

When we measure control rights concentration, we do not rely on a theoretical model as a basis for a concentration measure. Existing models of how large shareholders interact are based on probabilistic voting behavior under the

1. In general, votes at the annual meeting require a simple majority (50% plus one vote). However, changes to the charter (including equity issues) require approval of at least 75% (a "qualified majority") of the votes. Companies, in the charter, can set higher levels than the legal minimum of three-quarters of the votes, but few companies choose to do so.

assumption of one share, one vote. In addition, these theories are based on environments in which all shareholders are alike except that they have differing numbers of votes, e.g., the Shapley-Shubik Power Index (Shapley and Shubik, 1954) or the Banzhaf Index (Banzhaf, 1965, 1968). Leech (1988) and Leech and Leahy (1991) and the references cited therein provide further discussion. However, the German environment is much more complicated than these models. For example, it is not clear how to take proxy voting into account. There is also the issue of the identity of the shareholder, which can affect the shareholder's role and powers. For example, bank blockholders may not be the same as nonbank blockholders with the same number of votes. Indeed, this is something that we want to test for.

To measure the degree of control rights concentration in each firm we use a Herfindahl index (see, e.g., Demsetz and Lehn, 1985; and Cable, 1985). Recall that the Herfindahl index is defined as $H = \sum_{i=1}^{n} s_i$, where s_i $(i = 1, \ldots, n)$ is the fraction of stock owned by the agent i. If there are two agents, each holding 50% of the voting shares, H equals $0.5^2 + 0.5^2 = 0.5$. If there is a single agent who owns all the stock, H equals 1. The Herfindahl index is based on equity control rights, i.e., on control rights that emanate from ownership of voting stock, as discussed above. In particular, it does not include proxy votes. (Appendix B further discusses calculation of this index.)

13.2.4. Summary of German Equity Control Rights Structure

For the small samples, the control rights structure of each firm is measured with three variables: the banks' fraction of control rights from equity ownership (EB), the fraction of the votes that banks vote in proxy (VB), and the Herfindahl index of the concentration of control rights from equity ownership, H. The variable VB is measured relative to the actual presence at the annual meeting. The Herfindahl index comprises all blockholders, including banks, which enter H individually. For the large samples, VB is not available. With respect to the variables EB and H, it is important to note that, since the banks are included in the variable H, any effect we detect from the banks' control rights, EB, must be due to a channel that is different than that available to nonbank blockholders.

Proxy voting arises because German shares are generally bearer securities, and individual stockholders keep their shares at their bank. By agreement, German banks have the right to exercise proxy votes for these shareholders. Agreement is given in writing and lasts for 15 months. Shareholders can instruct the bank how to vote, if they wish, but this must be in writing. Banks do not, however, have unlimited power to vote shares held at the bank. Prior to the annual meeting, banks inform the shareholders they represent as to how they will vote at the meeting. If individual shareholders disagree with the bank, they can indicate how they want to vote by informing the bank (by mail). The bank must then adhere

to these instructions. Proxy-voting rights tend to be concentrated in the largest banks due to the fact that these banks happen to have an extensive network of branches. In the late 1970s, the largest six private (i.e., non-state-owned) banks controlled about three-quarters of the voting rights of dispersed shareholders (Krümmel, 1980). The Big Three banks (Deutsche Bank, Dresdner Bank, and Commerzbank) held just under half of the deposited shares in 1988 (Deutsche Bundesbank Monthly Report, April 1989).

Banks do not actively compete for proxy votes; banks with large networks of branches simply have many customers and these customers keep their shares at the bank without special instructions. From the banks' perspective, proxy voting is a passive byproduct of retail brokerage. In a similar vein, proxy voting might be viewed as the mirror image of the firm's shareholder structure, in particular its concentration of equity control rights, H. If this held, we would not expect proxy voting to be statistically significant in our empirical analysis.

Table 13.2, Panel A, provides the details of bank control rights from equity ownership, bank proxy voting, and the equity control rights of nonbank blockholders for the two small samples. Table 13.2, Panel B, covers the large samples. The tables show that equity ownership generally gives banks (as a group) control over far less than 25% of the votes. Also, proxy voting generally provides banks (as a group) with less than 25% of the votes at annual meetings. Thus, for the largest German firms (which compose our samples), control by banks, if it exists, does not appear to depend on the sheer number of votes. This point is reinforced by the fact that, in Germany, a large fraction of public companies have a single (nonbank) shareholder who holds at least 25% of the stock.

Our samples illustrate the importance of nonbank blockholders: 68 (264) out of 82 (283) firms in the small (large) 1975 sample have blockholders holding at least 25%; for the small (large) 1986 sample it is 46 (249) out of 56 (280). The pervasiveness of nonbank blockholders is not an aberration of our samples. Franks and Mayer (2000) study a sample of 171 German companies during the late 1980s and find that in 85% of these companies there is a single shareholder who holds at least 25%. Also, Edwards and Fischer (1994) report that "the vast majority of German AGs have a single shareholder who owns 25 percent or more of the voting capital" (p. 194). In contrast, a survey of exchange-listed firms in the U.S. in 1984 shows that only 20% of the firms have at least one nonofficer who owned 10% of firm stock; 13% of the firms are majority owned (Holderness and Sheehan, 1988). In the U.K. the proportion of public limited companies with a majority shareholder is also far smaller than in Germany (Edwards and Fischer, 1994).

13.2.5. Voting Restrictions

The voting rights of shareholders can be restricted by an AG's charter (articles of association) not to exceed some fraction of the total votes issued by the firm,

Table 13-2. BANK EQUITY CONTROL RIGHTS (CONTROL RIGHTS THAT EMANATE FROM BANKS' EQUITY OWNERSHIP), EB, BANKS' PROXY VOTING RIGHTS, VB, AND EQUITY CONTROL RIGHTS CONCENTRATION, H. THE HERFINDAHL INDEX OF CONCENTRATION OF EQUITY CONTROL RIGHTS, H, IS CALCULATED OVER ALL (BANK AND NONBANK) BLOCKHOLDERS, TREATING BANKS INDIVIDUALLY (I.E., NOT IN AN AGGREGATED FASHION). PANEL A: SMALL SAMPLES. PANEL B: LARGE SAMPLES. NOTE THAT THE LARGE SAMPLES DO NOT HAVE INFORMATION ON BANKS' PROXY VOTING AS MEASURED BY VB.

	1975 Sample	**1986 Sample**
Panel A		
(1) Bank equity control rights, EB Mean (median)	0.08 (0)	0.13 (0)
Standard deviation (min, max)	0.17 (0, 0.52)	0.31 (0, 2.03)
$0.00 \leq EB \leq 0.05$	61	40
$0.05 \leq EB < 0.1$	0	0
$0.1 \leq EB < 0.25$	4	3
$0.25 \leq EB < 0.50$	9	8
$0.50 \leq EB < 0.75$	8	4
$0.75 \leq EB \leq 1.00$	0	1
(2) Bank proxy voting rights, VB		
Mean (median)	0.21 (0.10)	0.23 (0.17)
Standard deviation (min, max)	0.28 (0, 0.90)	0.24 (0, 0.89)
$0.00 \leq VB \leq 0.05$	36	19
$0.05 \leq VB < 0.1$	5	4
$0.1 \leq VB < 0.25$	16	12
$0.25 \leq VB < 0.50$	12	14
$0.50 \leq VB < 0.75$	5	4
$0.75 \leq VB \leq 1.00$	8	3
(3) Equity control rights concentration, H		
Mean (median)	0.39 (0.26)	0.41 (0.28)
Standard deviation (min, max)	0.34 (0, 1)	0.34 (0, 1)
(4) Blockholders		
Number of firms with a block of		
at least 25% of control rights	68	46
at least 50% of control rights	38	25
at least 75% of control rights	20	15
(5) Sample size		
Total number of firms	82	56
Panel B		
(1) Bank equity control rights, EB		
Mean (median)	0.09 (0)	0.08 (0)
Standard deviation (min, max)	0.19 (0, 1.10)	0.20 (0, 2.03)
$0.00 \leq EB \leq 0.05$	208	223
$0.05 \leq EB < 0.1$	1	0
$0.1 \leq EB < 0.25$	7	15
$0.25 \leq EB < 0.50$	41	23
$0.50 \leq EB < 0.75$	21	17
$0.75 \leq EB \leq 1.00$	5	2

Table 13-2. (CONTINUED)

	1975 Sample	1986 Sample
(2) Equity control rights concentration, H		
Mean (median)	0.34 (0.26)	0.40 (0.32)
Standard deviation (min, max)	0.26 (0, 1)	0.29 (0, 1)
(3) Blockholders		
Number of firms with a block of		
at least 25% of control rights	264	249
at least 50% of control rights	163	172
at least 75% of control rights	61	79
(4) Sample size		
Total number of firms	283	280

regardless of the fraction of voting shares owned. Typical restrictions are 5% or 10%. Table 13.3 lists the firms and voting restrictions from our samples, also showing the year the restriction was adopted. (Most voting restrictions were adopted in the 1970s when Middle Eastern countries were looking for investment opportunities for their oil dollars and started to acquire stakes in German companies.) Clearly, this type of restriction constrains the power of block shareholders, including bank blockholders. Note, however, that banks's proxy voting of dispersed shareholders is not bound by this restriction, with Volkswagen AG being the only exception to this rule (Körber, 1989, pp. 97–98). These restrictions potentially make banks more powerful than nonbank shareholders and, consequently, it is not surprising that banks have supported these restrictions, though management has always initiated them (Edwards and Fischer, 1994).

Note that we do not expect the dummy variable for the presence of a voting restriction to be significant. If the firm's shareholder structure, along with bank proxy voting, explains the presence of a restriction, then it should have no separate, significant effect. As is possible with bank proxy voting, a voting restriction might simply be the mirror image of the firm's shareholder structure. This argument holds even in the case that the firm's shareholder structure (and the extent of proxy voting) changed in response to the adoption of a voting restriction.

13.2.6. Firm Performance Measures

For performance measures we use an accounting measure of profitability, the return on equity (ROE), and a market-based measure, the (log of the) market-to-book ratio (MTB). Accounting measures of firm performance have been widely used by other researchers, e.g., Demsetz and Lehn (1985), though in our case we rely on German accounting. Harris, Lang, and Möller (1994) find that the relation between 18-month stock returns and annual earnings for large

Table 13-3. VOTING RESTRICTIONS, BY COMPANY, BY TYPE, AND BY YEAR THEY WERE ADOPTED. VOTING RESTRICTIONS LIMIT THE NUMBER OF VOTES THAT EACH OWNER OF VOTING STOCK IS ALLOWED TO EXERCISE AT THE ANNUAL SHAREHOLDER MEETING. MOST VOTING RESTRICTIONS ARE BASED ON A FRACTION OF VOTES IN THE TOTAL VOTES ISSUED BY THE FIRM, WHILE OTHERS ARE BASED ON AN ABSOLUTE NUMBER OF VOTES. WITH THE EXCEPTION OF VOLKSWAGEN AG, VOTING RESTRICTIONS DO NOT APPLY TO VOTES THAT BANKS EXERCISE IN PROXY FOR SMALL SHAREHOLDERS. SOURCE: VERLAG HOPPENSTEDT, *Saling Aktienführer,* VARIOUS ISSUES, DARMSTADT.

Company with voting restriction	**Type of restriction**	**Year introduced**
Antriebstechnik G. Bauknecht AG	10%	1986
ASKO Deutsche Kaufhaus AG	5%	1977
AVA Allgemeine Handelsgesellschaft der Verbraucher AG	1%	1986
BASF AG	80 million Deutsche Marks of equity (face value)	1975
Bayer AG	5%	1975
Continental Gummiwerke AG	5%	1984
Hoesch AG	15%	1977
Industrie-Werke Karlsruhe Augsburg AG	10%	1985
Leifheit AG	10%	1985
Linde AG	10%	1973
Mannesmann AG	5%	1975
Rosenthal AG	5%	1986
Schering AG	12 million Deutsche Marks of equity (face value)	1973
Volkswagenwerk AG	2%/20%	1960/1970

German firms over the period 1982–1991 is basically the same as in the U.S. The market-to-book ratio is essentially Tobin's Q. While we do not construct estimates of the replacement costs of fixed assets or adjust for taxes, Perfect and Wiles (1994) show that these adjustments are not significant. For the large samples, the numbers of firms we use for the MTB and ROE regressions are the same. For the small samples, the number of firms in the MTB regressions is lower than in the ROE regressions because not all firms are traded.

Details on German accounting rules can be found in Coenenberg (1974, 1993) and Ordelheide and Pfaff (1994). We calculate the book value of equity as the sum of the face value of equity (including equity-like certificates), reserves, profits, and special reserves. The market- to-book ratio of equity, MTB, equals the 1976 (1987) year-end market value of equity (aggregated over all categories of stock) divided by the 1976 (1987) year-end book value of equity. (We linearly interpolate the book value of equity for the firms with other than calendar

fiscal years.) The return on equity, ROE, equals the surplus of the year 1976 (1987), divided by the book value of equity, averaged over fiscal year-ends 1976 and 1977 (1987 and 1988). Surplus of the year equals net profits plus payments to minority shareholders and the parent firm less any income obtained from the parent firm to cover losses. The book value of total assets is the sum of equity, provisions, and debt.

We also want to control for other exogenous characteristics of the sample firms that can affect performance. The following additional variables are included unless otherwise indicated: a codetermination dummy variable (Co) that equals one if there is equal representation, and zero otherwise; a voting restriction dummy variable (VR) that equals one if there is a voting restriction, and zero otherwise; a state ownership dummy variable (Go) that equals one if a majority of the voting shares are controlled by government entities, and zero otherwise; (log of) total assets (TA); and an industry dummy for industry j (ISIC j) based on the International Standard Industrial Classification (United Nations, 1990). We also include a dummy variable for the year 1986; this absorbs the change in the price deflator, which means that we do not have to deflate total assets.

13.3. GERMAN BANKS AND CORPORATE CONTROL: HYPOTHESES

In addition to measurement issues, there is the problem of specifying the link between firm performance and measures of equity control rights. The lack of theoretical guidance about this link motivates our empirical approach. In this section, we provide an overview of our approach and specify broad hypotheses to be examined.

13.3.1. Overview

We focus on how firm performance varies in cross-section as a function of (i) which fraction of the firm's votes is controlled by banks via equity ownership, EB, (ii) how much of the firm's equity banks vote in proxy, VB, (iii) the extent to which there are nonbank block shareholders, H, (iv) the degree to which the firm is subject to codetermination, Co, (v) the presence of voting restrictions, VR, and (vi) other factors (normalizing regressors) that capture characteristics of the firm that can affect performance.

We want to relate the ownership structure variables and the other independent variables to measures of firm performance. Let (EB_i, VB_i, H_i) be a vector of observations of the equity control variables of firm i; and let X_i be a (row) vector that represents Co_i, VR_i, and the observations from the set of normalizing

regressors. Let P_i be a measure of firm performance, either return on equity, ROE, or the (log of the) market-to-book ratio, MTB. For the reasons discussed above, we do not know how firm performance is affected by our three equity control variables, EB, VB, and H. Consequently, we initially investigate the performance of firm i ($i = 1, \ldots, n$) in the following semiparametric form:

$$P_i = X_i\beta + f(EB_i, VB_i, H_i) + \varepsilon_i, \qquad (\mathbf{13.1})$$

where $f(\cdot)$ is an unknown, possibly nonlinear, smooth function, but where the relation between X_i and performance is a (known) parametric function and ε_i is a mean-zero error term with variance σ^2. Based on specification tests using estimates of Eq. (13.1) we go on to parametric specifications.

The specification in Eq. (13.1) takes the equity ownership structure of firms as exogenous, reflecting the fact that we are studying an economy in which the stock market plays a much smaller role than in economies such as the U.S. or U.K. With a thin stock market, it is difficult for blockholders to assemble blocks in firms that they believe will do well in the future. Thus, we are proceeding under the view that Eq. (13.1) captures a potentially causal relation, e.g., bank block ownership causes firm performance according to the function specified. This view will be quite alien to those used to thinking about stock-market-based economies. To buttress our view, we document below that the equity ownership structures change little through time. There is little evidence that block positions respond to information about prospective firm performance. Eq. (13.1) also assumes that the firm's capital structure, the amount of bank borrowing, the amounts of retained earnings (i.e., dividend policy), and the composition of corporate boards are endogenous. These variables are at least partly determined by the same independent variables that determine P_i. (We discuss this further when we analyze the determinants of firm board composition.)

The specification in Eq. (13.1) treats banks in an aggregate fashion, that is, bank control rights from equity ownership and bank voting rights are each added up across banks. There are two reasons for this. First, empirically it is the case that there is usually a single bank that is the dominant bank equity holder for firms in which banks are important owners. This is related to the fact that equity ownership and proxy voting are concentrated in the largest banks. Second, the large banks, as a group, control a majority of votes at their own annual meetings (Gottschalk, 1988), strongly suggesting the possibility of collusion.

We now turn to discussing some hypotheses.

13.3.2. Bank Equity Ownership and Firm Performance

From Table 13.2 it might appear that bank equity holding is unimportant because nonbank blockholders are much more pervasive than bank

blockholders. Bank control rights from equity ownership, in general, seem low. But the conclusion that banks are not important would be premature. First, as discussed above, there can be voting restrictions in place, allowing banks to outvote large nonbank blockholders using proxy votes. Second, and perhaps more importantly, the power to exercise corporate control is not only a function of the allocation of formal control rights in the form of votes. Banks can have superior power and information that they use to their advantage even if their control rights are low in number and there is a large nonbank blockholder. Banks can also have superior information by virtue of the lending relationship (Elsas and Krahnen, 1998). In addition, as mentioned above, banks have power because they guard access to capital markets.

If banks can affect firm performance by virtue of having control rights that emanate from equity ownership, then there are three possibilities for how firm performance could be altered. First, if there is a coincidence of interests between banks and other shareholders, then banks can be benign or even improve performance. While banks' control-rights-derived power can give them the ability to expropriate from other shareholders, banks might not have the economic incentive to behave this way. Bank cash-flow rights can be highly correlated with control rights from equity ownership, the effect emphasized by Jensen and Meckling (1976), resulting in a coincidence of interests. In fact, while nonbank blockholders can improve firm performance to the extent that they hold control rights and cash-flow rights, banks are better able to improve firm performance than nonbank blockholders. In other words, what we will call the "coincidence-of-interests hypothesis" states that over the entire range of bank ownership of voting stock, the relationship between firm performance and the fraction of bank equity control rights is upward sloping, ceteris paribus.

A second possibility, maintained by strong critics of universal banking, is that the interests of bank equity holders and other shareholders are in opposition to each other, no matter how many votes the banks control via share ownership. Banks act in their own private interests to the detriment of other shareholders. For this hypothesis to hold, banks must have private benefits at stake, so that when the banks' block increases, they use the additional control rights to extract more private benefits. For example, by virtue of their dual role as lenders and equity holders, and to the extent that capital markets are not a very competitive financing option, banks can behave as monopolists, using their power to extract profit from the firm at the expense of firm performance. The view that German banks act as monopolists to the detriment of firm value is a long-standing criticism. Even the Deutsche Bundesbank disingenuously notes:

> When enterprises are deciding on which financing methods to adopt, the advice of their principal bankers may sometimes be to take up new loans, because the share issue which might be to the advantage of the enterprise is

> not rated so highly by the bank; however, definite statements in this regard can neither be made nor proved. (Monthly Report, April 1984, p. 15)

For example, monopoly profits can be extracted by increased borrowing from the bank, possibly at monopoly interest rates.

Finally, the relation between firm performance and the fraction of voting rights that banks control via equity ownership could be downward sloping over some initial range of bank equity ownership, and then upward sloping, ceteris paribus. That is, the bank faces a tradeoff between its private benefits and the value of its shares depending on its ability to extract private benefits. Such a tradeoff can depend on the size of the bank's equity stake. Holding other variables constant, a bank can face a conflict of interest over a low range of low equity holding, but not when its equity holding is high. In the case of such a conflict of interest, the relation between firm performance and bank equity control rights is nonlinear: firm performance can initially decline with an increase in the amount of control that is associated with an increase in bank equity ownership; when bank equity ownership and the corresponding fraction of equity control rights are large, firm performance rises with bank equity ownership.

The three descriptions of possible relations between firm performance and bank control rights from equity ownership are those that hold whenever there is a potentially informed insider blockholder in a system with one share, one vote. These are the hypotheses explored for U.S. managers' stockholdings by, for example, Morck, Shleifer, and Vishny (1988), and McConnell and Servaes (1990), and for banks by Gorton and Rosen (1995). The only difference here is that the bank can be potentially more informed and more powerful than managers and the bank can have more private benefits at stake. More important, however, are the interactions of the other characteristics of the governance system with bank control rights that emanate from ownership in voting stock. We now turn to these other characteristics.

13.3.3. Proxy Voting and Conflicts of Interest

A clear (at least formal) break between the alignment of control rights and cash-flow rights is in the ability of German banks to vote shares in proxy. This raises the prospect that banks vote in their private interests rather than in the interests of shareholders. Clearly, proxy-voting power is potentially important because the votes of dispersed shareholders are concentrated in banks. These votes can be used when important decisions are made at the general meeting. In particular, membership on the supervisory board is determined by elections at the general meeting. (By law, AGs must hold a shareholder meeting at least once a year.) Also, as discussed above, blockholders' voting power can be limited by voting

restrictions, which increases the importance of bank proxy voting. Thus, proxy voting by banks, which creates a concentration of voting power, would seem to generate the clearest possibility of a conflict of interest and, for this reason, has been very controversial in Germany.

Proxy voting gives banks control rights in excess of cash-flow rights. If proxy voting affects firm performance, then the possibilities for how banks use their proxy votes are the same as for the banks' control rights from equity ownership, which we discussed above. In the case of a coincidence of interests between banks and other shareholders or, in the opposite case, when interests are always in opposition to each other, an appropriate measure of bank control rights would be one for which proxy-voting rights add to the control rights from equity ownership. But how the excess control rights are used might depend on the level of the bank's cash-flow rights. That is, it could be that with low amounts of equity ownership the bank uses the proxy votes to enforce decisions in its private interests, while at high levels of equity holdings the bank uses proxy votes to maximize the value of the firm. In this case, there would be a critical value of bank control rights from equity ownership such that performance is increasing in bank proxy rights above this level and decreasing below it. In other words, there would be a critical fraction of bank equity control rights, EB^*, such that, holding everything else constant, $\partial P/\partial VB > 0$ for $EB > EB^*$ and $\partial P/\partial V\,B < 0$ for $EB \leq EB^*$.

Alternatively, bank proxy-voting rights might simply be the flip side of the firm's equity control structure, in particular, its concentration, H. In this case, proxy voting is endogenous and therefore should have no impact of its own.

13.3.4. Nonbank Block Shareholders

In stock market economies, outside block shareholders are often viewed as monitors of firm management because, by virtue of the size of their stake in the firm, they have an incentive to actively oversee management. Implicit in this view is a close link between control rights and cash-flow rights. In stock market economies, dispersed small shareholders can face free-rider problems in monitoring firm management if monitoring is costly (Grossman and Hart, 1980; Shleifer and Vishny, 1986). The empirical evidence for the U.S., while somewhat mixed, appears to support the importance of large shareholders in increasing firm value.[2] The potential behavior of banks, outlined above, can interact with the behavior of nonbank blockholders, but there are several possibilities for this interaction.

Since, as mentioned above, a very high percentage of the largest quoted German companies have a single shareholder owning at least 25% of the

2. See Demsetz and Lehn (1985), Mikkelson and Ruback (1985), Holderness and Sheehan (1988), Barclay and Holderness (1991), and Zeckhauser and Pound (1990).

shares, the monitoring role of blockholders might be very important in Germany and might explain why hostile takeovers are not necessary and hence are rare. Nonbank blockholders might be so powerful that they not only monitor firms' management but also monitor banks, preventing banks from falling prey to their conflicts of interest. On one hand, nonbank blockholders can behave as insiders, reducing firm performance over a range of low equity holdings by extracting private benefits but then improving firm performance when their equity holdings are high. Perhaps banks attempt to monitor the deleterious behavior of these blockholders. On the other hand, banks can collude with large blockholders. Basically, a number of (nonlinear) interactions with the bank ownership of voting rights and proxy voting are plausible. These considerations suggest the importance of controlling for the entire equity voting structure of the firm in attempting to detect the effects of banks on performance and further emphasize the importance of the econometric specification issue.

13.3.5. Equity Voting Restrictions

Voting restrictions delink control rights and cash-flow rights at the restriction point. Such voting restrictions potentially increase the power of bank proxy voting. Voting restrictions can also limit the size of nonbank blockholders and hence increase the power of banks, whether it emanates from votes or from other sources. As discussed below, however, it is likely that voting restrictions are endogenous, that is, they are a function of the equity ownership structure and hence should have no separate effect.

13.3.6. Codetermination

Corporate governance and firm performance in Germany can be influenced by the fact that, under German law, employees of large firms are allocated (voting) seats on the supervisory board. In Germany, the board system consists of the supervisory board (Aufsichtsrat) and the management board (Vorstand). The role of the supervisory board is to oversee the management board; it has the power to hire and fire, set compensation, regularly meet with management, and so on. Basically, the management board runs the day-to-day operations and is responsible to the supervisory board. According to German codetermination laws, employees must constitute either one-half or one-third of the firm's supervisory board, depending on the size of the firm. Some firms are not required to have employees on the supervisory board. Codetermination implies that a sizable fraction of the nonexecutive directors cannot be appointed by shareholders, even if a single shareholder would effectively be in control otherwise. This uncouples control rights and cash-flow rights, which makes codetermination

potentially important to the extent that the supervisory board controls the important decision-making of the firm.

There are three different forms of codetermination in Germany (see Wiedemann, 1980, Gorton and Schmid, 1998, for details). First, there is codetermination in the coal and steel industry (Montan-codetermination). It was introduced in 1951 and requires equal representation between employees and shareholders on the supervisory board. There is also a so-called neutral member on the supervisory board, to break ties. Second, the Codetermination Act of 1976 extended equal representation (with modifications) to all other industries, leaving Montan-codetermination in place. This law requires that if the corporation has regularly more than 2,000 employees, then the employees must elect one-half of the supervisory board members. Typically, about one-third of the employee representatives are members of the works council while the remainder consists of external trade union representatives. Even though half the seats go to workers, representation under the 1976 Codetermination Act is not quite equal because the chair, appointed by the shareholders, has an extra vote. Also, at least one employee representative must be elected from the senior managers. Third, under the Works Constitution Act of 1952, one-third employee representation is required of companies with 500 to 2,000 employees.

The effects of codetermination on the performance of a firm are potentially quite complicated. It could be that codetermination affects only the distribution of the firms' cash flows, but not its amount. That is, employees use their power on the supervisory board to bargain for a greater share of the firm's cash flows, but have no other effects. Whether employees have enough power to do this depends on whether other institutions, perhaps banks, can counteract such power. This is an empirical question. But codetermination can have other effects as well. If employees are risk averse and have firm-specific human capital at stake, then they can use their power on the supervisory board to alter the firm's investment and operating decisions in favor of reducing idiosyncratic firm risk. Furthermore, it could simply be the case that employees make poor decisions and hence reduce firm performance. Gorton and Schmid (1998) empirically explore many of these issues. Here, we limit ourselves to the question of whether codetermination is detrimental to firm value by taking account of cross-section variation in codetermination. Note that we account for the 1976 Codetermination Act in our 1975 samples because our firm performance measures are taken from the fiscal year 1977.

13.3.7. The Exogeneity of the Equity Ownership Structure

The specification in Eq. (13.1) assumes that the equity ownership structure and, in particular, bank blockholding, is exogenous or at least predetermined with respect to firm performance. When the stock market is not the dominant

institution for organizing the savings-investment process, it is difficult for agents to alter their portfolios. By definition, illiquidity is a central feature of a bank-based economy and the exogeneity of the ownership structure flows from this fact. It is precisely this relative illiquidity that makes bank-based economies different from stock-market-based economies. But exactly how illiquid are the stock markets in bank-based economies? Our main focus, however, is not on empirically examining the relative liquidity of the German stock market (though that seems like an interesting question). Our interest is whether banks are active equity portfolio managers, buying stock in undervalued firms and selling blocks in overvalued firms. To address this question with respect to banks we examine how banks acquire their equity positions and how these positions change through time. The basic point is that German banks are not actively managing equity portfolios, which would imply the existence of a liquid stock market.

Typically, banks acquire blocks of shares as byproducts of banking relationships; blocks are purchased from families or during distress. The Deutsche Bundesbank reports:

> German banks originally acquired part of their shareholdings... via special transactions or through "rescue operations" for enterprises which had got into liquidity difficulties. Portfolio considerations alone never tip the scales when banks are contemplating the purchase of equities. (Monthly Report of the Deutsche Bundesbank, April 1984, p. 16)

"Special transactions" refer to purchases of blocks from family owners who are selling out.[3] For details on block trades in Germany see Franks and Mayer (2000).

Besides the illiquidity of the stock market, there are strong tax incentives for not selling blocks of equity that, possibly due to active monitoring of bank blockholders, have appreciated over time. Capital gains are not taxed before being realized through sale. Capital gains from block sales are subject to the full corporate tax rate, which gives blockholders an incentive to hold on to their equity stakes. (At the end of the year 1999, the German government revealed plans to lower the tax rate that applies to realized capital gains from block trades, in an attempt to lower the transaction costs of equity control changes and encourage corporate restructuring.)

3. Studienkommission (1979, p. 87) reports that 559 of the 662 bank equity participations observed at the end of 1974 (they sent out a questionnaire and only considered cases where 10% or more was held) were acquired after the year 1948. Most of these holdings were acquired after 1960. Herrhausen (1987, p. 107, Table 3) presents some information on why banks hold equity. He considers 20 acquisitions of the ten largest private banks that took place in the period 1976–1986. Only seven of these companies were traded at the stock exchange at this time. The reasons mentioned by these banks were: long-term investment (six cases), short-term investment (five cases), support of medium sized companies which are weakly endowed with capital (five cases), credit rescue measure (one case), anti-takeover measure (one case), and other reasons (two cases).

As a result of the illiquidity of the stock market and the tax incentives, it is not surprising that German equity ownership structures tend not to change much through time. In particular, the block ownership of firms by banks is persistent. Table 13.4 details the ownership shares in some large companies by the Big Three, Deutsche Bank (Panel A), Dresdner Bank (Panel B), and Commerzbank (Panel C). The table covers the period 1972–1990. (Recall that our samples are drawn from 1975 and 1986.) While there is some change in equity ownership, the main feature is the persistence of block size over the period.

The illiquidity of equity, and bank blocks in particular, is potentially important for the German system of corporate governance. A number of researchers, including Maug (1998), Kahn and Winton (1995), and Admati, Pfleiderer, and Zechner (1994), explore the choice of block size and the behavior of the blockholders, viewing blockholders as (possibly risk-averse) monitors of firms (also see Bhide, 1993). A blockholder can monitor management and in the process become privately informed about the firm. Such a blockholder faces a decision concerning whether to trade on this private information or continue as a blockholder. In an economy with a liquid stock market, a blockholder faces a number of these types of decisions. But in an economy where the stock market is less liquid, or simply illiquid, such tradeoffs do not occur. Blockholders, especially banks, can be forced to try to maintain or improve the value of blocks, as monitors of the firm's management, because the alternative of selling the blocks is not available.

13.4. ECONOMETRIC METHODOLOGY

As discussed above, a number of hypotheses involve nonlinearities between firm performance, bank control rights from equity ownership, EB, bank proxy voting, VB, and equity control rights concentration, H, while other hypotheses imply monotonic relations. Since the shape of Eq. (13.1) is critical to our investigation, our approach is to start by using a semiparametric estimation procedure to search for nonlinearities. We want to allow the data to dictate the functional form so we avoid having to arbitrarily specify a parametric form for Eq. (13.1). We test for the appropriate semiparametric specification (i.e., "window size," as discussed below) but also include some parametric functions as potential candidates. Our strategy is to try to impose structure on Eq. (13.1) in a step-by-step fashion, starting from as little structure as possible and proceeding by letting the data guide us, possibly to a parametric form.

13.4.1. Semiparametric Estimation: Overview

Eq. (13.1) consists of a parametric part (*the term* $X\beta$) and a nonparametric part, the function $f(\cdot)$. We want to allow full generality as to the possible shape of

Table 13-4. Share ownership in nonfinancial firms of the largest German banks, the Big Three, for the period 1972–1990. The table is in favor of our assumption that in Germany, holdings of large blocks (by banks in particular) show a sufficient degree of persistence to be treated as an exogenous variable in our empirical analysis. The addendum i stands for indirect ownership as defined by Böhm (1992), our data source. Note that his definition of indirect ownership complies only roughly with our concept of ultimate owners. Panel A: Deutsche Bank AG. Panel B: Dresdner Bank AG. Panel C: Commerzbank AG. Source: Böhm (1992).

Year	**1972**	**1975**	**1978**	**1980**	**1982**	**1984**	**1986**	**1988**	**1990**
Panel A									
Stock Corporations									
AEG AG	0	0	0	0	0	>5	16 i	16 i	22.5 i
Bergmann Elektrizitätswerke AG	>25	>25	>25	>25	>25	>25	36.5	36.5	36.5
Continental AG	10 i	10 i	10	10	10	10	10	10	10
Daimler Benz AG	>25	>25	28.5	28.5	28.5	28.5	28.1	28.2	28.1
Hapag Lloyd AG	>25	>25	>25	>25	>25	>25	>25	12.5	12.5
Philipp Holzmann AG	>25	>25	>25	>25	>35	>35	>25	35.4	30
Horten AG	18.8 i	18.8 i	18.8 i	18.8 l	18.8 i	18.8 i	18.8 i	18.8 i	18.8 i
Karstadt AG	>25	>25	>25	>25	>25	>25	>25	>25	>25
Klöckner-Humboldt-Deutz AG	0	0	0	0	0	0	0	41.5 i	41.1 i
Klöckner Werke AG	0	0	0	0	0	0	0	19.6 i	0
Linde AG	0	0	0	0	0	0	0	10	10
Metallgesellschaft AG	8.3 i	8.3 i	8.3 i	13.1 i	8.8 i	8.8 i	10.6 i	10.7 i	10.1 i
Nixdorf AG	0	0	0	0	25	0	0	0	0
VEW AG	6.3 i	6.3 i	6.3 i	6.3 i	6.3 i l	6.3 i	6.3 i	6.3 i	6.3 i
Firms of other legal forms									
MBB GmbH	0	0	0	0	0	0	0	0	17.7
MTU GmbH	0	0	14.3 i	14.3 i	14.3 i	14.3 i	28.1 i	28.2 i	28 i

Table 13-4. (CONTINUED)

Year	1972	1975	1978	1980	1982	1984	1986	1988	1990
Panel B									
Stock Corporations									
AEG AG	0	0	0	0	0	>5	0	0	0
Bayerische Motoren Werke AG	0	0	0	0	0	5 i	5 i	5 i	5 i
Bilfinger und Berger AG	>50	44	44	>25	>25	>25	>25	>25	25
Continental AG	0	0	0	0	0	0	0	0	7.7
Degussa AG	10 i	10 i	10 i	10 i	10 i	10 i	10 i	10 i	10 i
FAG Kugelfischer KGaA	0	0	0	0	0	>10	0	0	0
Hapag Lloyd AG	>25	>25	>25	>25	>25	>25	>25	12.5	12.5
Kaufhof AG	>25	>25	>25	>25	9	9	9	0	0
Metallgesellschaft AG	>25	>25	25	30	33	16.5 i	18 i	23.1 i	23.3 i
Firms of other legal forms									
MBB GmbH	0	0	0	0	0	5 i	5 i	5 i	5 i
Panel C									
Stock Corporations									
FAG Kugelfischer AG	0	0	0	0	0	>10	0	0	0
Hochtief AG	>25	>25	25	12.5 i	12.5 i	12.5 i	12.5 i	12.5 i	12.5 i
Philipp Holzmann AG	0	0	0	0	5	>7.5 i	>7.5 i	5 i	>7.5 i

Year	**1972**	**1975**	**1978**	**1980**	**1982**	**1984**	**1986**	**1988**	**1990**
Horten AG	6.3 i	6.3 i	7.3 i	6.3 i	6.3 i	6.3 i	6.3 i	6.3	6.3
Karstadt AG	>25	>25	>25	>25	>25	>25	>25	>25	>25
Kaufhof AG	>25	>25	>25	>25	0	0	0	0	0
Linde AG	0	0	0	10	10	10	10	10	10
MAN AG	0	7.5 i	7.5 i	7.5 i	6.2 i	6.2 i	7.5 i	7.5 i	7.5 i
Sachs AG	0	0	25	25	>25	>25	>35	0	0
Firms of other legal forms									
Thyssen AG	0	0	0	0	0	0	5 i	5 i	5 i

$f(\cdot)$. Estimation of Eq. (13.1) and inference are complicated by the combination of the parametric component with the nonparametric, smooth component. We follow a procedure proposed by Speckman (1988). The basic approach is to purge each component of dependence on the other component and then apply ordinary least squares to the parametric part and a (linear) smoother to the nonparametric part. Consequently, we start by defining

$$X^* = (I - K)X \tag{13.2}$$

and

$$P^* = (I - K)P. \tag{13.3}$$

These are the variables X and P, "adjusted" for dependence on EB, VB, and H, via the smoother matrix K. (I is the identity matrix.) Then β is estimated by

$$\widehat{\beta} = \left(X^{*'}X^*\right)^{-1} X^{*'}P^* \tag{13.4}$$

and the estimate of the nonparametric part reads

$$\widehat{f} = K \cdot \left(P - X\widehat{\beta}\right). \tag{13.5}$$

With regard to the choice of K, we use (quadratic) locally weighted regression, LOESS (Cleveland and Devlin, 1988; Müller, 1987; Stute, 1984; and Cleveland, 1979). The advantage of LOESS over kernel methods is that it can handle multidimensional smoothing with fairly small data sets. LOESS cannot only account for possible nonlinear effects the variables EB, VB, and H, might have in isolation. LOESS can also control for possible interactions among these three explanatory variables as they affect firm performance. Such interaction effects would, for example, be observed if banks fell prey to their conflicts of interest.

13.4.2. Specification Testing: The M-Statistic

While locally weighted regression does not require a functional form to be specified, it does require that a smoothing parameter, g, be chosen. Based on Mallows' (1973) C_p criterion, Cleveland and Devlin (1988) developed a method that offers some guidance in the choice of this smoothing parameter. We outline this procedure in the following.

Let z_i be the triplet $\{EB_i, VB_i, H_i\}$ for firm i. The function $f(\cdot)$ at point z_i is estimated uses the q nearest neighbors of this data point. The smoothing parameter g is the fraction of the q nearest neighbors in the number of observations in the sample, i.e., $g = q/n$. Thus, the estimate, $\hat{f}_g(z_i)$ depends on g, as does it mean squared error.

The expected mean squared error summed over z_i, $i = 1, \ldots, n$, and divided by σ^2 is

$$M_g = \frac{E \sum_{i=1}^{n} \left(\hat{f}_g (z_i) - f(z_i) \right)^2}{\sigma^2}. \tag{13.6}$$

Eq. (13.6) shows how the choice of the smoothing parameter, g, trades off variance of the estimator against bias. For a sufficiently small value of the smoothing parameter, $g = g_o$, the bias of $\hat{f}_g (z_i)$ is negligible, resulting in a nearly unbiased estimate of σ^2. Let s^2 an estimate of σ^2 for the smoothing parameter g_o. Also, let

$$B_g = \frac{e_g' e_g}{s^2} - tr \left(I - K_g \right)' \left(I - K_g \right) \tag{13.7}$$

and

$$V_g = tr\, K_g' K_g, \tag{13.8}$$

where e_g is the vector of residuals obtained when the smoothing parameter g is employed. The subscript g on K indicates the dependence of the smoother on g. The expected mean squared error, M_g, can be estimated by

$$\hat{M}_g = \hat{B}_g + V_g. \tag{13.9}$$

$\hat{B}_g$ is the contribution of bias to the estimated mean squared error and V_g is the contribution of variance. When $\hat{f}_g (\cdot)$ is a nearly unbiased estimate, then the expected value of $\hat{B}_g$ is nearly zero, so the expected value of $\hat{M}_g$ is nearly V_g. As g increases, bias is introduced, and $\hat{B}_g$ has a positive expected value, so the expected value of $\hat{M}_g$ exceeds V_g.

V_g is called the equivalent number of parameters of the fit by analogy with the Mallows (1973) C_p statistic. The equivalent number of parameters decreases as the smoothing parameter, g, increases, i.e., more structure is imposed. Cleveland and Devlin (1988) show that the distribution of $\hat{M}_g$, the M-statistic, is (approximately) an F distribution under the assumption of no bias. Cleveland and Devlin (1988) describe the degrees of freedom and Cleveland, Devlin, and Grosse (1988) describe Monte Carlo studies of the approximation. Using this result, we can calculate the distribution of the M-statistic for any $g \geq g_o$ under the null hypothesis no bias. We will convey this information with a graph of $\hat{M}_g$ against V_g, the equivalent number of parameters. The plots will also show the 90% confidence intervals.

We plot the M-statistic for our semiparametric specification over a range of smoothing parameters, g, and for two parametric specifications. We are interested in specifications for which bias is negligible. The M-statistic does not directly test one specification against another (i.e., it is not directional), but this serves our purposes because we are not testing against a particular

alternative hypothesis. Whang and Andrews (1993) discuss directional tests in the semiparametric context.

13.5. THE EFFECTS OF BANKS ON FIRM PERFORMANCE

In this section we estimate the performance relation in Eq. (13.1) and draw inferences about some of the hypotheses outlined above. We first address the issue of the shape of Eq. (13.1). If we detect nonlinearities, then, depending on the details of the nonlinearity, this could be evidence in favor of one of the conflicts-of-interest hypotheses. That is, there could be ranges of equity control rights over which there is a detectable effect on performance of the uncoupling of cash-flow rights and control rights. If there are such nonlinearities, it will rule out the straightforward monotonic hypotheses that banks have either coincident or opposing interests over all ranges of the firms' multidimensional control rights structures.

Based on the results concerning the shape of Eq. (13.1), the analysis proceeds by estimating a parametric specification, addressing the question of which equity control rights variables, EB, VB, or H, affect firm performance. We then analyze changes in German corporate governance between 1975 and 1986 and compare our results to Cable (1985).

13.5.1. The Shape of the Performance-Ownership Structure Relation with Proxy Voting

We start by focusing on the small samples because they contain proxy-voting measures. The issue of conflicts of interest seems most important here and therefore, the issue of nonlinearities is most critical.

Figure 13.2 is an M-plot for the market-to-book ratio for the small 1975 sample from g $g = 0.65$ to $g = 1.0$, with steps of 0.05. (Since our data sets are small, we start out with a fairly high smoothing parameter to avoid the problem of overfitting.) In the figure, the rightmost $\times$-symbol is for $g = 0.65$, which increases from right to left (because V_g decreases) until we come to the leftmost $\times$-symbol. We also include two parametric specifications: quadratic (i.e., including squared and cross-terms of EB, VB, and H) and linear (without such terms). The leftmost box symbol is the linear specification; the other box is the quadratic specification. In the figure, the upward-sloping line is $\hat{M}_g = V_g$, assuming no bias for the lowest value of the smoothing parameter, $g = 0.65$. The vertical lines are 90% confidence intervals. The figure shows that the quadratic and the linear parametric specifications are unbiased for the (log of the) market-to-book ratio, MTB. Figure 13.3 shows the M-plot for the return on equity, ROE, for the small 1975 sample. Again, both quadratic and linear

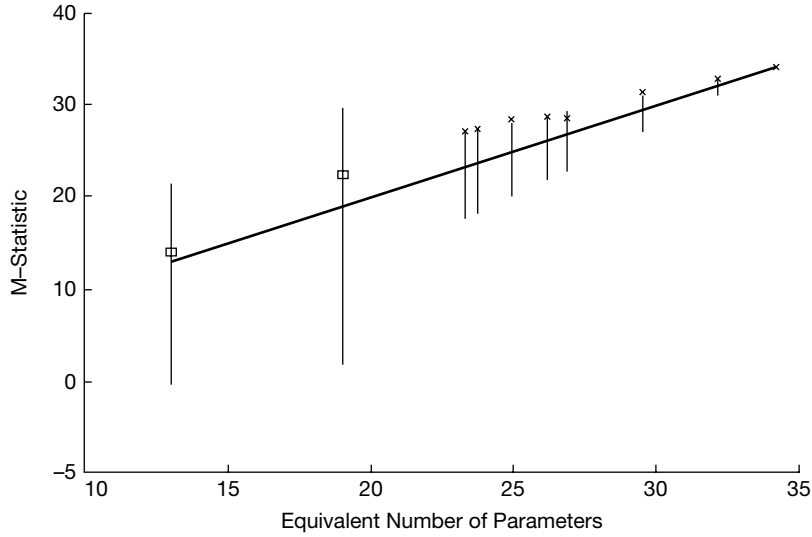


Figure 13.2 M-plot for the small 1975 sample when firm performance is measured by the (log of the) market-to-book ratio of equity, MTB. The upward-sloping line is drawn under the assumption that the bias in the semiparametric estimation is negligible for the lowest value of the smoothing parameter we applied, $g = 0.65$. The $\times$-symbols represent alternative values for the smoothing parameter. The M-statistic and the equivalent number of parameters that comes with the lowest smoothing parameter is represented by the rightmost $\times$-symbol. The smoothing parameter increases in steps of 0.05 from right to left. The two box symbols represent parametric specifications; the right box stands for a quadratic least-squares specification (which includes squared and cross-terms of EB, VB, and H), while the left box is a linear least-squares specification (i.e., one without such terms). The vertical lines are 90% confidence intervals around the null hypothesis that the specification in question delivers unbiased estimates of the unknown functional form.

parametric specifications are acceptable in terms of bias. This conclusion means that (for the small 1975 sample) we cannot reject the null that there are no nonlinearities; hypotheses implying such nonlinearities are not supported by the data because the relation is monotonic in all control rights variables, EB, VB, and H.

We now turn to the small 1986 sample. Figures 13.4 and 13.5 show the M-plots for this sample. Because this sample is smaller than the 1975 sample, we start with a larger smoothing parameter. The plot begins with $g = 0.75$ and increases to $g = 1.0$ by steps of 0.05. The symbols are as in the previous plots. Note that the symbols for the quadratic and the linear parametric specifications are within the 90% confidence interval. As for the 1975 sample, this means that the data do not support the nonlinear hypotheses for the 1986 sample.

The specification tests of the large samples give similar results. (The M-plots are omitted.) Note that the large samples do not have proxy-voting data. Thus, the nonparametric part of Eq. (13.1) has two dimensions only (EB and H).

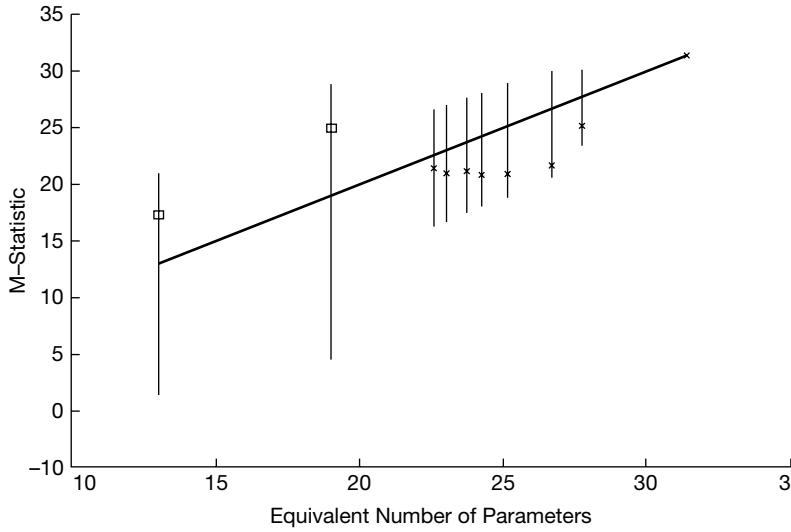


Figure 13.3 M-plot for the small 1975 sample when firm performance is measured by the return on equity, ROE. The upward-sloping line is drawn under the assumption that the bias in the semiparametric estimation is negligible for the lowest value of the smoothing parameter we applied, $g = 0.65$. The $\times$-symbols represent alternative values for the smoothing parameter. The M-statistic and the equivalent number of parameters that comes with the lowest smoothing parameter is represented by the rightmost $\times$-symbol. The smoothing parameter increases in steps of 0.05 from right to left. The two box symbols represent parametric specifications; the right box stands for a quadratic least-squares specification (which includes squared and cross-terms of EB, VB, and H), while the left box is a linear least-squares specification (i.e., one without such terms). The vertical lines are 90% confidence intervals around the null hypothesis that the specification in question delivers unbiased estimates of the unknown functional form.

We find that for both performance measures, linear parametric specifications are acceptable in terms of bias. This is our first important finding. The remaining questions are whether banks affect firm performance and, if so, whether the interests of banks are in opposition to or coincident with those of other shareholders. We try to answer these questions by examining the linear parametric specification.

13.5.2. Are the Conflicts of Interest Between Banks and Other Shareholders?

We now present least squares performance regressions for each sample (small and large). We pool the two years, 1975 and 1986, in a single regression and test for differences across years.

Table 13.5 shows the results for MTB for the small sample and Table 13.6 shows the results for ROE for the small sample. From these tables we learn that (i) when MTB is the performance measure, firm performance increases

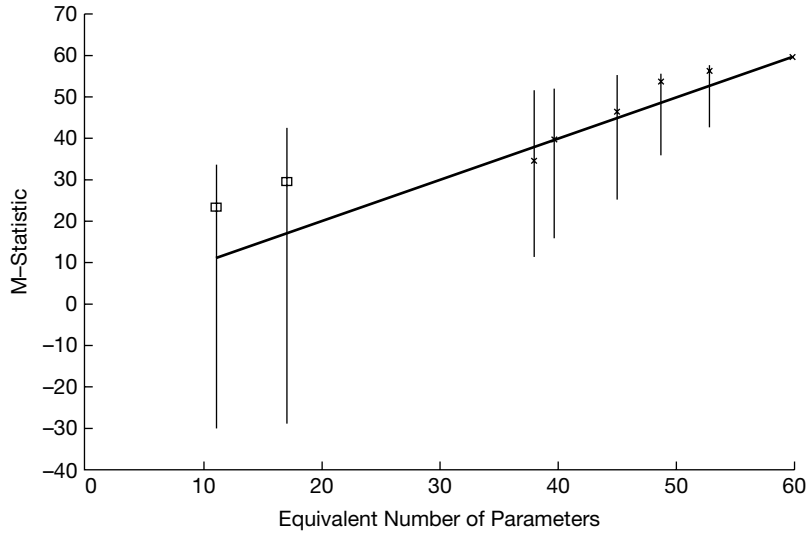


Figure 13.4 M-plot for the small 1986 sample when firm performance is measured by the (log of the) market-to-book ratio of equity, MTB. The upward-sloping line is drawn under the assumption that the bias in the semiparametric estimation is negligible for the lowest value of the smoothing parameter we applied, $g = 0.75$. The $\times$-symbols represent alternative values for the smoothing parameter. The M-statistic and the equivalent number of parameters that comes with the lowest smoothing parameter is represented by the rightmost $\times$-symbol. The smoothing parameter increases in steps of 0.05 from right to left. The two box symbols represent parametric specifications; the right box stands for a quadratic least-squares specification (which includes squared and cross-terms of EB and H), while the left box is a linear least-squares specification (i.e., one without such terms). The vertical lines are 90% confidence intervals around the null hypothesis that the specification in question delivers unbiased estimates of the unknown functional form.

as a function of the banks' control rights from equity ownership, EB; (ii) firm performance is not related to bank proxy voting as measured by VB; (iii) firm performance is positively related to concentration of control rights from equity ownership, H; (iv) when ROE is the performance measure, firm performance decreases with codetermination.

The results using the large samples are displayed in Tables 13.7 and 13.8. The large samples do not contain the proxy voting variable, VB. Table 13.7 shows the large sample results for the MTB ratio and Table 13.8 contains the results for ROE. Firm performance is increasing in the banks' control rights from equity holdings, EB, when the MTB ratio is the performance measure. Nonbank blockholding also improves MTB and codetermination causes MTB to decline. The results using ROE as a performance measure are essentially noise.

Overall, we can summarize the results as follows. The first result is that banks affect firm performance beyond the effects they would have if they were nonbank blockholders. An increase of the banks' control rights from equity ownership by one percentage point (i.e., 100 basis points) changes the market-to-book ratio

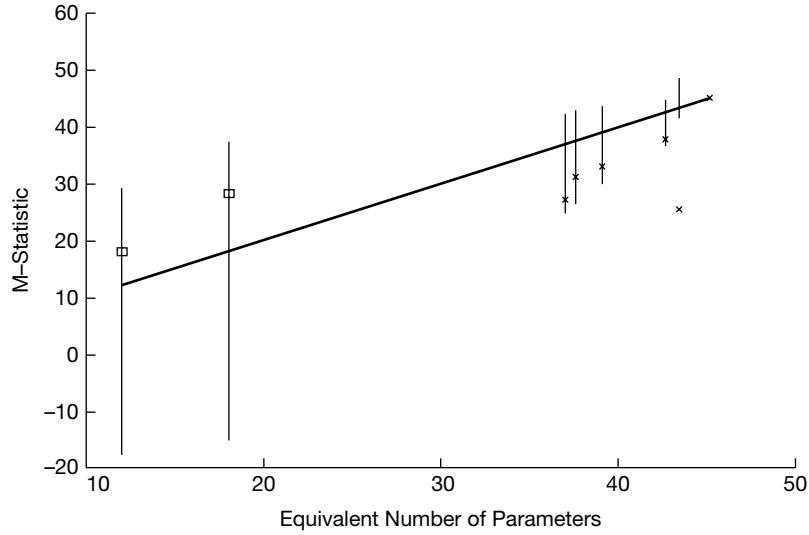


Figure 13.5 M-plot for the small 1986 sample when firm performance is measured by the return on equity, ROE. The upward-sloping line is drawn under the assumption that the bias in the semiparametric estimation is negligible for the lowest value of the smoothing parameter we applied, $g = 0.75$. The $\times$-symbols represent alternative values for the smoothing parameter. The M-statistic and the equivalent number of parameters that comes with the lowest smoothing parameter is represented by the rightmost $\times$-symbol. The smoothing parameter increases in steps of 0.05 from right to left. The two box symbols represent parametric specifications; the right box stands for a quadratic least-squares specification (which includes squared and cross-terms of EB and H), while the left box is a linear least-squares specification (i.e., one without such terms). The vertical lines are 90% confidence intervals around the null hypothesis that the specification in question delivers unbiased estimates of the unknown functional form.

of the firm by 0.23% in the small sample and by 0.41% in the large sample. The power of the banks cannot be due to the fact that they are blockholders because banks are included in the Herfindahl index of concentration of control rights, H. Thus, banks appear to be special in positively affecting firm performance.

Second, banks' proxy voting, VB, does not affect firm performance. In particular, there do not appear to be any conflicts of interest between banks' use of proxy voting and shareholders' interests. A possible reason for the statistical insignificance of VB can be that proxy voting is a mirror image of the firm's shareholder structure, which is sufficiently controlled for by EB and H.

Third, the concentration of control rights from equity ownership, H, is important in improving firm performance.

Finally, codetermination reduces firm performance. If β is the regression coefficient of a dummy variable in a semi-logarithmic regression equation, then $100\left(e^{\beta} - 1\right)$ equals the percentage change of the dependent variable caused by

Table 13-5. LEAST-SQUARES ESTIMATES OF THE INFLUENCE OF CODETERMINATION, Co, BANKS' EQUITY CONTROL RIGHTS, EB, BANKS' PROXY VOTING, VB, AND CONCENTRATION OF EQUITY CONTROL RIGHTS, H, ON FIRM PERFORMANCE. FIRM PERFORMANCE IS MEASURED BY THE (LOG OF THE) MARKET-TO-BOOK VALUE OF EQUITY, MTB. THE DATASET POOLS OBSERVATIONS FROM THE SMALL 1975 AND 1986 SAMPLES. NORMALIZING REGRESSORS INCLUDE A DUMMY VARIABLE FOR VOTING RESTRICTIONS, VR, A DUMMY VARIABLE FOR GOVERNMENT-CONTROLLED FIRMS, Go, (THE LOG OF) TOTAL ASSETS AS A MEASURE FOR FIRM SIZE, TA, A DUMMY VARIABLE FOR THE OBSERVATIONS FROM THE 1986 SAMPLE, DUMMY VARIABLES FOR INDUSTRY CLASSIFICATION, AND A CONSTANT TERM. STANDARD ERRORS ARE CORRECTED FOR HETEROSKEDASTICITY FOLLOWING WHITE (1980).

Independent variable	**Coefficient**	***t*-value**
Co	-9.92×10^{-2}	–0.75
EB	2.30×10^{-1}	1.82*
VB	1.29×10^{-1}	0.61
H	5.28×10^{-1}	2.08**
VR	2.47×10^{-3}	0.02
Go	-3.88×10^{-1}	–1.49
TA	-1.20×10^{-2}	–0.28
Dummy 1986	2.90×10^{-2}	0.27
ISIC C	-6.66×10^{-1}	–2.74***
ISIC D	1.02×10^{-1}	0.63
ISIC E	-7.73×10^{-2}	–0.56
ISIC F	6.58×10^{-2}	0.57
ISIC G	-9.86×10^{-3}	–0.05
Constant	4.87×10^{-1}	0.54
R^2 adj.	0.06	
Wald-statistic	44.6***	
Number of observations	96	

*Significant at 10% level (two-tailed *t*-tests).
**Significant at 5% level (two-tailed *t*-tests).
***Significant at 1% level (two-tailed *t*-tests).

a change of the dummy variable from zero to one (see Halvorsen and Palmquist, 1990). A change in the codetermination dummy variable from zero to one (i.e., a switch from no codetermination or one-third codetermination to equal representation) reduces the market-to-book ratio by 15.9% in the large sample; ROE is reduced by 3.25 basis points in the small sample. (The other cases have insignificant coefficients.)

13.5.3. Changes Between 1975 and 1986

We now ask whether the effects of the firm's control rights structure on firm performance change significantly between 1975 and 1986. To examine this issue

Table 13-6. Least-squares estimates of the influence of codetermination, Co, banks' equity control rights, EB, banks' proxy voting, VB, and concentration of equity control rights, H, on firm performance. Firm performance is measured by the return on equity, ROE. The dataset pools observations from the small 1975 and 1986 samples. Normalizing regressors include a dummy variable for voting restrictions, VR, a dummy variable for government-controlled firms, Go, (the log of) total assets as a measure for firm size, TA, a dummy variable for the observations from the 1986 sample, dummy variables for industry classification, and a constant term. Standard errors are corrected for heteroskedasticity following White (1980).

Independent variable	**Coefficient**	***t*-value**
Co	-3.25×10^{-2}	-3.04^{***}
EB	2.24×10^{-3}	0.19
VB	5.58×10^{-3}	0.28
H	5.50×10^{-2}	2.96^{***}
VR	1.95×10^{-2}	1.49
Go	-2.47×10^{-2}	−1.60
TA	1.38×10^{-2}	3.17^{***}
Dummy 1986	-6.16×10^{-3}	−0.54
ISIC C	-3.35×10^{-2}	-2.58^{***}
ISIC D	1.59×10^{-2}	1.29
ISIC E	1.01×10^{-2}	0.76
ISIC F	-1.06×10^{-2}	−1.03
ISIC G	2.05×10^{-3}	0.14
Constant	-2.57×10^{-1}	-2.82^{***}
R^2 adj.	0.13	
Wald-statistic	36.1^{***}	
Number of observations	138	

***Significant at 1% level (two-tailed *t*-tests).

we test whether the coefficients on the control rights variables, EB, VB, and H, are significantly different between these two dates. Note that the large sample does not contain bank proxy voting, as measured by VB. We present results from Wald-tests that are based on a heteroskedasticity-consistent variance-covariance matrix as proposed by White (1980).

For the small sample, the results are as follows. When the performance measure is the MTB ratio, $\chi^2(3) = 5.120$ and $p = 0.163$, and when ROE is the performance measure, $\chi^2(3) = 0.201$ and $p = 0.977$. In the large sample, when the performance measure is the MTB ratio, $\chi^2(2) = 2.409$ and $p = 0.300$, and when ROE is the performance measure, $\chi^2(2) = 2.319$ and $p = 0.314$. Thus, there are no significant differences in the influence of the control rights structure between the years 1975 and 1986.

Table 13-7. LEAST-SQUARES ESTIMATES OF THE INFLUENCE OF CODETERMINATION, CO, BANKS' EQUITY CONTROL RIGHTS, EB, AND CONCENTRATION OF EQUITY CONTROL RIGHTS, H, ON FIRM PERFORMANCE. FIRM PERFORMANCE IS MEASURED BY THE (LOG OF THE) MARKET-TO-BOOK VALUE OF EQUITY, MTB. THE DATASET POOLS OBSERVATIONS FROM THE LARGE 1975 AND 1986 SAMPLES. NORMALIZING REGRESSORS INCLUDE A DUMMY VARIABLE FOR VOTING RESTRICTIONS, VR, A DUMMY VARIABLE FOR GOVERNMENT-CONTROLLED FIRMS, GO, (THE LOG OF) TOTAL ASSETS AS A MEASURE FOR FIRM SIZE, TA, A DUMMY VARIABLE FOR THE OBSERVATIONS FROM THE 1986 SAMPLE, DUMMY VARIABLES FOR INDUSTRY CLASSIFICATION, AND A CONSTANT TERM. STANDARD ERRORS ARE CORRECTED FOR HETEROSKEDASTICITY FOLLOWING WHITE (1980).

Independent variable	**Coefficient**	***t*-value**
Co	-1.74×10^{-1}	–2.19**
EB	4.09×10^{-1}	3.70***
H	3.30×10^{-1}	3.70***
VR	4.29×10^{-2}	0.35
Go	-2.72×10^{-1}	–2.12**
TA	-4.43×10^{-2}	–2.76***
Dummy 1986	2.29×10^{-1}	4.12***
ISIC A	1.10	3 24***
ISIC C	1.71×10^{-1}	0.60
ISIC D	2.99×10^{-1}	4.71***
ISIC E	1.66×10^{-1}	1.13
ISIC F	-7.97×10^{-2}	–0.80
ISIC G	4.31×10^{-1}	3.06***
ISIC H	-6.99×10^{-2}	–0.98
ISIC I	4.74×10^{-1}	1.84*
ISIC J	-1.97×10^{-1}	–1.75*
Constant	1.08	3.54***
R^2 adj.	0.12	
Wald-statistic	145***	
Number of observations	563	

*Significant at 10% level (two-tailed *t*-tests).
**Significant at 5% level (two-tailed *t*-tests).
***Significant at 1% level (two-tailed *t*-tests).

13.5.4. Comparison of the Results to Cable (1985)

Cable (1985) is the only previous study of the effects of German bank relationships on German firms' performance. Cable uses a subset (48 AGs) of our sample for 1975. He averages other variables over the period 1968–1972. Cable's dependent variable, a performance measure, is the ratio of the after tax income of equity to total assets of the firm. While Cable estimates many models, the most general includes (i) the square of each bank's voting fraction, (ii) a dummy

Table 13-8. Least-squares estimates of the influence of codetermination, Co, banks' equity control rights, EB, and concentration of equity control rights, H, on firm performance. Firm performance is measured by the return on equity, ROE. The dataset pools observations from the large 1975 and 1986 samples. Normalizing regressors include a dummy variable for voting restrictions, VR, a dummy variable for government-controlled firms, Go, (the log of) total assets as a measure for firm size, TA, a dummy variable for the observations from the 1986 sample, dummy variables for industry classification, and a constant term. Standard errors are corrected for heteroskedasticity following White (1980).

Independent variable	**Coefficient**	***t*-value**
Co	1.92×10^{-3}	0.20
EB	3.91×10^{-3}	0.49
H	1.31×10^{-2}	0.83
VR	2.07×10^{-2}	1.84*
Go	2.81×10^{-3}	0.28
TA	7.82×10^{-4}	0.50
Dummy 1986	-1.84×10^{-2}	–2.45**
ISIC A	9.74×10^{-4}	0.05
ISIC C	2.81×10^{-2}	1.20
ISIC D	1.82×10^{-2}	3.14***
ISIC E	6.57×10^{-4}	0.06
ISIC F	3.19×10^{-4}	0.06
ISIC G	9.47×10^{-3}	1.57
ISIC H	3.50×10^{-6}	0.00
ISIC I	-9.93×10^{-3}	–1.49
ISIC J	2.25×10^{-2}	1.33
Constant	2.29×10^{-2}	0.77
R^2 adj.	0.002	
Wald-statistic	50.9***	
Number of observations	563	

*Significant at 10% level (two-tailed *t*-tests).
**Significant at 5% level (two-tailed *t*-tests).
***Significant at 1% level (two-tailed *t*-tests).

variable for each of the three largest banks that equals one if the bank has supervisory board seats, (iii) the ratio of total bank borrowing to total debt, (iv) a Herfindahl index of the top 20 nonbank shareholders, and (v) normalization variables.

There are a number of important differences between Cable's approach and ours. First, calculation of Cable's performance measure is debatable because it divides the income of the equity holders by total assets (i.e., the numerator of return on equity is divided by the denominator of the return on assets). Second,

our view is that board membership and bank borrowing are endogenous. (Cable includes the ratio of total bank borrowing to total debt as an independent variable but it would seem to depend on the ownership variables, which he also includes.) Thirdly, Cable does not differentiate between the votes that banks cast in proxy and the votes that they hold as owners of firm equity (he includes the sum of the two).

Although it is hard to interpret Cable's results, his own conclusion is that there is a significant positive impact on firm performance from interaction with banks. Edwards and Fischer argue that "Cable's study provides considerably more support for the view that what is distinctive about German AGs is their typically concentrated share ownership, which means that there are incentives for large shareholders to monitor management carefully, and so improve profitability" (p. 226). Our results are not in agreement with this interpretation. Instead, we support Cable's own conclusion because we showed that banks are special; they affect firm performance in a way that cannot be attributed simply to their role as blockholders.

13.6. BANKS AND THE SUPERVISORY BOARD

The ability to influence firm performance could be related to membership on the firm's supervisory board, the board that has important power in running the firm. In this section, we examine bank representation on the firms' supervisory boards.

Bank representation on supervisory boards has been almost as controversial as bank proxy-voting power. The Monopolkommission (1980) finds that commercial bank representatives accounted for 9.8% of all supervisory board members of the 100 largest AGs in 1978 and were represented on 61 of the top 100 boards. The largest three banks held 94 of the 145 bank representatives. In 1974, banks held seats on the supervisory boards of 59 out of the 74 officially quoted large companies (Studienkommission, 1979; Krümmel, 1980).

We did not use the supervisory board representation of banks as an explanatory variable in our regressions, because the power that comes from board representation is power that is "derived" from equity control rights as measured by EB, VB, and Herf. However, we are interested in knowing whether equity control rights translate into supervisory board membership. It is important to stress that this is not necessary for firm performance to be affected by a bank relationship, though we are interested in whether it is a channel of influence.

For our analysis, the dependent variable is the number of seats held by banks divided by the number of seats allocated to shareholder representatives. (No honorary board members are taken into account.) Appendix A provides detail on the data sources. We use the same independent variables as before except that

we do not include the industry dummies (because they are, as a group, not statistically significant). Also, for this analysis we use a Herfindahl index that *excludes* banks (HNB), with the fraction of equity owned by nonbanks (as a group) normalized to unity. Previously, we wanted to identify bank power as distinct from the power of nonbank blockholders, so we included banks in the Herfindahl index, H. For the analysis of board seats, we do not include banks in the index, because banks and nonbank blockholders can be in competition for seats. Also, we included slope dummies for the influence of (the log of) total assets, instead of relying on the intercept dummy to pick up changes in the price deflator. This allows us to interpret the intercept dummy in a meaningful way as a measure of change in the autonomous fraction of board seats occupied by banks.

The dependent variable is a fraction that is bounded at zero and has indivisibilities, which are particularly relevant for its numerator because the number of seats occupied by banks is an integer. Thus, the dependent variable is censored. We therefore estimate a Tobit model. A drawback here is that the size of the board varies among the sample firms and thus the indivisibilities might not have the same effect for all the firms.

The results for the pooled sample are shown in Table 13.9. In both 1975 and 1986, bank control rights from equity ownership are significant in determining the fraction of supervisory board seats that banks hold. A χ^2 test for the joint significance of the intercept dummy variable and the slope dummies for the EB, VB, and HNB gives $\chi^2(4) = 13.99$ and $p = 0.007$, indicating that there is a statistically significant structural break between 1975 and 1986.

The regressions presented by Edwards and Fischer (1994, pp. 198–210) use the same underlying data set on supervisory board membership as we do and as Cable (1985) did for the 1975 sample. However, the dependent variable and the sample in our analysis will differ from Edwards and Fischer in ways that turn out to be important. First, Edwards and Fischer restrict their sample to those stock corporations (51 firms) for which banks cast more than 5% of the votes at the annual meetings of 1975 (votes from equity ownership plus proxy votes). (This is because that is the way the Monopolkommission provided this information.) However, the remaining firms have negligible values for EB and VB, mostly because these firms are closely held. For this reason we do not restrict ourselves to those 51 companies that Edwards and Fischer analyze. Another issue with the Edwards and Fischer results is that these authors use the absolute numbers of seats (held by banks) as the endogenous variable. However, the total number of seats on the supervisory board in their sample of 51 companies varies between three (for Triumph International AG) and 21 (for August Thyssen-Hütte AG, for example). (See Verlag Hoppenstedt, *Handbuch der deutschen Aktiengesellschaften,* 1974/75 and 1975/76 issues, Darmstadt.)

Table 13-9. TOBIT ESTIMATION OF THE INFLUENCE OF CODETERMINATION (CO), BANKS' EQUITY CONTROL RIGHTS (EB), AND CONCENTRATION OF NONBANK SHAREHOLDERS' CONTROL RIGHTS (HNB), ON THE FRACTION OF (VOTING) SUPERVISORY BOARD SEATS HELD BY BANKS. THE FRACTION OF THE SUPERVISORY BOARD SEATS OCCUPIED BY BANKS WAS MEASURED RELATIVE TO THE NUMBER OF SUPERVISORY BOARD SEATS THAT ARE ASSIGNED TO SHAREHOLDER REPRESENTATIVES (AS OPPOSED TO THOSE THAT ARE ASSIGNED TO EMPLOYEE REPRESENTATIVES). THE DATASET POOLS OBSERVATIONS FROM THE SMALL 1975 AND 1986 SAMPLES. NORMALIZING REGRESSORS INCLUDE A DUMMY VARIABLE FOR VOTING RESTRICTIONS (VR), A DUMMY VARIABLE FOR GOVERNMENT-CONTROLLED FIRMS (GO), (THE LOG OF) TOTAL ASSETS AS A MEASURE FOR FIRM SIZE (TA), A DUMMY VARIABLE FOR THE OBSERVATIONS FROM THE 1986 SAMPLE, DUMMY VARIABLES FOR INDUSTRY CLASSIFICATION, AND A CONSTANT TERM. THE VARIANCE-COVARIANCE MATRIX WAS ESTIMATED FOLLOWING EICKER (1967) AND WHITE (1980).

Independent variable	**Coefficient**	***t*-value**
Co	-1.12×10^{-2}	–0.18
EB 1975	6.10×10^{-1}	4.20***
EB 1986	1.78×10^{-1}	3 41***
VB 1975	1.66×10^{-1}	2.09**
VB 1986	1.96×10^{-1}	1.93*
HNB 1975	-1.02×10^{-1}	–1.28
HNB 1986	8.06×10^{-3}	0.14
VR	4.87×10^{-3}	0.14
Go	7.94×10^{-2}	0.52
TA 1975	4.96×10^{-3}	0.23
TA 1986	-3.57×10^{-2}	–1.85*
D 1986	8.77×10^{-1}	1.43
Constant	5.36×10^{-2}	0.12
χ^2 (structural break)	14.0***	
χ^2 (nonconstant regressors)	48.3***	
Number of positive observations	116	
Number of observations	138	

*Significant at 10% level (two-tailed *t*-tests).
**Significant at 5% level (two-tailed *t*-tests).
***Significant at 1% level (two-tailed *t*-tests).

13.7. DISCUSSION OF THE RESULTS

In a stock-market-based economy, corporate governance can occur via assembling blocks to take over or influence managers when this intervention is valuable. In a bank-based economy, there is no market for corporate control. Instead, banks are heavily involved in corporate governance. Dow and Gorton (1997) argue that bank-based economies can, in theory, be just as efficient as stock market economies. While our results are consistent with this general notion,

there are many important missing details. Our results pose many questions for further research. In this section we briefly discuss some of these questions.

The two most important questions are interrelated. First, what is the source of bank power that makes it possible for banks to improve the value of firms? Second, what are the incentives that induce banks to use their power to improve firm performance, as opposed to extracting private benefits to the detriment of firm performance? Our results are consistent with the view that bank blockholders, having acquired a block of stock from a family or as a result of distress, have an incentive to monitor the firm if the stock market is illiquid. Basically, when the stock market is illiquid the bank blockholder can only sell at a large loss (Bhide, 1993). This creates an incentive to maintain a close relationship with the firm. In fact, the illiquidity commits the bank to monitor. This argument applies to all blockholders, while our results go further to distinguish banks from other blockholders in their ability to affect performance; banks are more powerful than nonbank blockholders because they improve firm performance beyond what nonbank blockholders can achieve. For example, Bethel, Liebeskind, and Opler (1998) find that in the U.S., "activist" blockholders (e.g., raiders) are more effective than institutional blockholders in causing value-increasing changes at firms. It is not simply a matter of counting up the number of votes held by a blockholder. Thus, the important question is: What is special about banks compared to nonbank blockholders? One possibility is that banks have more power than nonbank blockholders because banks have the credible threat of cutting off external finance. Just as banks cannot feasibly sell their blocks, without liquid capital markets, firms have no outside option for financing and must rely on their banks. The absence of a deep stock market forces banks and firms into a symbiotic relationship that can substitute for disciplining via takeovers. Another (nonmutually exclusive) possibility is that banks have better information, and possibly superior expertise, relative to other blockholders.

Why do banks improve firm performance? Why do they not act in their private interests? One answer concerns the possible positive correlation between bank control rights from equity ownership and bank ownership of cash-flow rights. To the extent that banks own cash-flow rights they have a financial incentive to improve the performance of firms and will use their power to this end (Jensen and Meckling, 1976; La Porta, Lopez-de-Silanes, Shleifer, and Vishny, 1999). Bank ownership of control rights and cash-flow rights could be positively correlated despite the institutional features, such as codetermination, voting restrictions, pyramiding, cross-shareholdings and stocks with multiple votes, that act to uncouple them. The fact that banks have cash-flow rights in the form of loans, as well as equity claims, might be important in this regard.

Another (nonmutually exclusive) explanation for the behavior of banks concerns the issue of who monitors the banks. In a purely formal sense, Diamond's argument about "monitoring the monitor" might apply in Germany, but

certainly the depositors of a bank would not mind if the bank extracted private benefits from client firms if they could benefit from this. However, in Germany, banks may be treated as quasi-public institutions, a view that is perhaps consistent with the degree of public scrutiny they receive. It is also consistent with the view of Allen and Gale (1997), who present a model of (German) banking that relies on a sort of social compact to set up and maintain the banking system with a fixed rate of interest on deposits (i.e., it does not vary across the business cycle). In their overlapping generations framework, some generations have an incentive to renege on this compact but, for unexplained reasons, do not. Clearly, these issues remain unresolved.

Another question for further research concerns proxy voting. If banks improve performance with respect to their own holdings, why do they not use their proxy power to further improve firm performance? There are several possible explanations for this result. First, banks simply may not need this additional power. Second, were banks to use their power overtly (even if for the good) they might face social sanctions. Finally, bank power is limited by the ability of individuals to tell banks how to vote. If individuals felt this were necessary to do, they might prefer to deposit their stock with another bank. Competitive pressure thus may limit bank power.

13.8. CONCLUSION

Little is known about corporate governance in economies in which the stock market is not a central institution. In economies with stock markets, the link between control rights and cash-flow rights is more direct and, consequently, can be the basis for takeovers as the ultimate form of governance. Poorly run firms can be taken over by a raider who buys shares in the stock market. Because a share purchase is the purchase of a bundle of cash-flow rights and control rights, the raider will have an incentive and the power to improve the value of the firm. In economies with small or nonexistent stock markets, banks appear to be very important. The concentration of effective, if not formal, power in banks is in contrast to the workings of stock market economies. Our investigation focuses on the extent to which a bank relationship in Germany affects firm performance when the mechanism of takeovers is absent and banks appear powerful.

What happens in economies in which the stock market is not so liquid and listings are few? In Germany, several institutional features, aside from the small stock market, suggest that the link between cash-flow rights and control rights is somewhat uncoupled. In particular, with respect to corporate governance, place Germany has the following notable features: (i) bank equity ownership, (ii) proxy voting by banks, (iii) high concentration of equity ownership, and (iv) codetermination. We empirically investigate whether these features interact in ways that provide a role for banks to positively affect the performance of firms.

When doing that we take into account (i) voting restrictions, (ii) pyramiding, (iii) cross-shareholdings, and (iv) stocks with multiple votes.

We find evidence supporting the notion that banks are an important part of the corporate governance mechanism in Germany. Firm performance, measured by the market-to-book value of equity, improves to the extent that banks have control rights from equity ownership. During the periods we investigate, banks do not extract private value to the detriment of firm performance. We find no evidence of conflicts of interest between banks and other shareholders. In particular, we find no evidence that banks use proxy voting to further their own private interests or, indeed, that proxy voting is used at all. It appears, then, that corporate governance mechanisms that are different from those that operate in stock-market-based economies can be effective. Clearly, however, many questions remain to be studied.

APPENDIX A: DATA SOURCES

A.1. The 1975 Samples

The small 1975 sample is constructed from the list of the top 100 stock corporations *(Aktiengesellschaften)* of the year 1974, published in Monopolkommission (1978). The criteria for choosing the firms are described in Monopolkommission (1977).

Of these 100 companies, we drop 18 companies: three firms were joint ventures of nonprofit cooperatives; two firms published their unconsolidated reports according to the accounting rules of banks; two firms were Kommanditgesellschaften auf Aktien, a hybrid ownership form between a stock corporation and a partnership; two firms published only consolidated financial statements; two firms were in the process of restructuring (one of them after a change in ownership); one firm did not publish an annual report; five firms were in financial distress; and, finally, for one firm we could not determine the ownership.

The accounting data on each firm and information on voting restrictions are from *Handbuch der deutschen Aktiengesellschaften* and from *Saling Aktienführer,* Verlag Hoppenstedt, Darmstadt, various issues. Information on bank proxy voting (for the small sample) comes from reports on annual shareholder meetings that took place in 1975, published in Monopolkommission (1978). Information on equity ownership structure was collected for the year 1975; it is from Monopolkommission (1977), from *Handbuch der deutschen Aktiengesellschaften,* various issues, and from *Saling Aktienführer 1976.*

The large 1975 sample consists of all nonfinancial firms listed in *Saling Aktienführer 1976.* This volume covers all stock corporations traded in the first

market segment (amtlicher Handel) or the second market segment (geregelter Freiverkehr) at any German stock exchange at the end of September 1975. Of 425 firms, we drop 142: seven were Kommanditgesellschaften auf Aktien; two firms published their unconsolidated reports according to bank guidelines; seven were nonprofit companies (six public transportation firms and one real estate firm); five were firms in the process of liquidation; one firm did not publish unconsolidated financial statements; 37 were real estate firms (most of which are "zombies," i.e., they have liquidated their production facilities); five were financial holding shells (firms whose main business is to hold equity stakes in other firms without serving as concern headquarters); 31 were firms in financial distress; 23 were delisted from the exchange within the next two years (i.e., within the period of time we measure firm performance); and 24 firms were missing information on ownership structure. We classify a firm as financially distressed if its equity's book value falls short of 110% of its equity's face value, i.e., the book value was lower than the face value plus the mandatory reserves, and the company is not a startup firm.

A.2. The 1986 Samples

The small 1986 sample is drawn from the list of the 100 largest (by sales, based on consolidated figures) German manufacturing firms (of all legal forms) published on October 3, 1986 by the *Frankfurter Allgemeine Zeitung.* Thus, unlike the 1975 sample, the 1986 sample contains no retailers, transport, or media companies. We follow Böhm (1992) in using this list because he is our main source for the bank proxy voting data. The list contains 65 stock companies. Of these we drop nine companies: one firm was in the process of restructuring (after a change in ownership); three firms were Kommanditgesellschaften auf Aktien; and five firms were in financial distress.

Company data, including equity ownership, are again from *Handbuch der deutschen Aktiengesellschaften* and from *Saling Aktienführer,* various issues. Information on the equity ownership structure dates from 1986. Information on bank proxy voting comes from three sources: Gottschalk (1988), Böhm (1992), and our own survey of annual shareholder-meeting reports (procured from commercial registers in the province where the company is chartered), which corrected and supplemented the other sources. Proxy voting data are based on the attendance lists of annual meetings that took place in calendar year 1986. (The 1986 report of the annual meeting of Siemens AG was not available at the commercial register in Munich; we thus used the 1985 report.)

The large 1986 sample consists of all nonfinancial firms listed in *Saling Aktienführer1987* (published in 1986). Again, this volume covers all stock corporations

traded in the first (amtlicher Handel) and second market segment (geregelter Markt) at any German stock exchange at the end of September 1986. Of 432 firms, we dropped 152: four were Kommanditgesellschaften auf Aktien; seven were nonprofit companies; two firms were in the process of liquidation; one firm was in the process of restructuring; one firm was a target of a battle over a minority shareholder position (which heavily affected its stock value); eight firms filed for bankruptcy within the next two years (the period of time we measure firm performance); 52 were real estate firms (again, most of which are "zombies"); seven were financial holding shells; 54 were firms in financial distress; and 16 firms were delisted from the exchange within the next two years (i.e., within the period of time we measure firm performance).

Table 13.A1 describes the industry classification of the firms included in the small samples. Table 13.A2 describes the industry classification of the firms included in the large samples.

A.3. Supervisory Board Membership Data

For the 1975 sample, data on board representation are taken (as in Edwards and Fischer, 1994, pp. 198–210) from Monopolkommission (1978). The 1986 data on board representation are taken from Bohm (1992, pp. 257–262) and from *Handbuch der deutschen Aktiengesellschaften,* various issues.

A.4. Additional Notes

(1) Both small samples are drawn based on size measures from consolidated reports. We have no control over this because we want to use the available proxy voting data that had already been collected based on these samples. However, we use unconsolidated financial statements. Since German firms can choose among several consolidation methods, their consolidated financial statements are poorly comparable over time and in cross-section. Also, since consolidation includes companies that are only partially owned by the firm in question, the analysis of unconsolidated reports has the advantage of providing a close link between equity ownership and firm performance.

(2) In both samples, and for the analysis of supervisory boards, Kreditanstalt-für Wiederaufbau and Bayerische Landesanstalt für Aufbaufinanzierung are not treated as banks because they are government-controlled special purpose banks (for reconstruction and development). The first one is a federal institution and the latter one is a Bavarian bank. In our sample they are treated as government institutions.

Table 13-A1. Distribution of firms in the small 1975 and 1986 samples by International Standard Industrial Classification (ISIC) as published by United Nations (1990). The classification was undertaken by the authors because there is no publicly available official industry classification of the corporations in our sample.

Number of firms 1975/1986	ISIC category	Industrial classification
5/1	C	Mining and Quarrying
54/38	D	Manufacturing
9/10	E	Electricity, Gas and Water Supply
6/5	F	Construction
6/2	G	Wholesale and Retail Trade; Repair of Motor Vehicles, Motorcycles and Personal and Household Goods
2/0	—	Not Classified (Highly Diversified)
Total: 82/56		

Table 13-A2. Distribution of firms in the large 1975 and 1986 samples by International Standard Industrial Classification (ISIC) as published by United Nations (1990). The classification was undertaken by the authors because there is no publicly available official industry classification of the corporations in our sample.

Number of firms 1975/1986	ISIC category	Industrial classification
2/2	A	Agriculture, Hunting and Forestry
3/2	C	Mining and Quarrying
217/218	D	Manufacturing
26/23	E	Electricity, Gas and Water Supply
8/7	F	Construction
9/16	G	Wholesale and Retail Trade; Repair of Motor Vehicles, Motorcycles and Personal and Household Goods
1/1	H	Hotels and Restaurants
11/9	I	Transport, Storage and Communications
2/0	K	Real Estate, Renting and Business Activities
4/2	—	Not Classified (Highly Diversified)
Total 283/280		

APPENDIX B: EQUITY CONTROL RIGHTS AND EQUITY OWNERSHIP STRUCTURE

This appendix explains some of the assumptions and methods of calculation concerning the ownership structure of firms' control rights and also the calculation of the Herfindahl indices. The equity ownership data are not always

detailed enough to obtain a complete picture of the equity control rights ownership structure. To calculate the Herfindahl index, we need to know, in addition to the details of bank equity holdings, the distribution of shares across nonbank blockholders and the percentage of shares that are dispersed. Tables 13.1 and 13.2 show some of the details of bank and nonbank ownership of voting rights, but to calculate the index we use data that are further disaggregated. In some cases, however, it is necessary to make some assumptions to complete the picture of equity ownership in order to calculate the index. We first explain these assumptions here. We then provide more information concerning how control rights from equity ownership are calculated, by providing some examples of the more complicated ownership structures.

B.1. Assumptions Concerning Equity Ownership

In some cases, vote holdings are reported as greater than 25%, greater than 50%, greater than 75%, less than 25%, etc. In these cases, we adopt the following conventions (unless other information can make determination of the holdings more precise): we set "greater than 25%" equal to 26%; we set "greater than 50%" equal to 51%; etc. The reported inequalities refer to cutoff points that are relevant for control purposes as discussed in Section 13.2. In other words, if x is the fraction of shares held by the particular blockholder, "greater than 25%" means $0.5 > x \geq 0.25$.

We assume that the banks vote all dispersed holdings if no other information can make this more precise. The bank proxy voting is originally reported as a percentage of votes in attendance at the annual shareholder meeting. Bank proxy voting at the annual meeting is taken to be dispersed shareholders' votes (though on rare occasions this is not true). We assume that shareholders that do not show up at the annual meeting are dispersed. (Note that this assumption applies only to calculation of the Herfindahl index and not to the fraction of bank proxy votes.)

An example will show how the aforementioned assumptions are used. For simplicity, we assume that for all blockholders in this example, the fraction of control rights equals the fraction of voting stock owned (i.e., there are no pyramids, cross-shareholdings or stocks with multiple votes). Let B_1 be the fraction of shares voted by blockholder 1 and B_2 the fraction voted by blockholder 2, etc. Suppose the data are that $EB = 0, B_1 > 0.25$, and $B_2 = 0.1$, and the rest are dispersed. The problem is that we do not know the exact size of B_1's holdings. If we have no other information, we assume $B_1 = 0.26$. However, from the proxy-voting fraction that banks vote at the annual meeting we can calculate VB under the assumption that the banks vote all dispersed shares. Then we obtain $B_1 = 1 - a \times VB - 0.1$, with a being the fraction of votes present at the annual meeting.

B.2. Control Rights When Equity Ownership Is Complex

We give two examples of complex equity ownership structures, and how we calculated control rights in these cases. The first example is a case of a pyramid with direct and indirect holdings, shown in figure 13B.1. Following our principle of defining control rights based on votes, the graph displays ownership as fractions of votes, which is not necessarily identical to the fractions of equity from which these votes are derived. On September 30, 1986, Energieversorgung Ostbayern AG was owned by Bayernwerk AG (a nonfinancial firm) with more than 50% of the shares, Energiebeteiligungs-Gesellschaft mbH (a financial holding shell) with more than 25% of the shares, and the State of Bavaria with 1.7%. As shown in the figure, the complications are first that 75% of Energiebeteiligungs-Gesellschaft mbH is owned by CONTIGAS Deutsche Energie-AG, a publicly traded utility, and 25% by Bayernwerk AG, which is also a utility but is not publicly traded. In addition, Bayernwerk owns 54% of CONTIGAS and 35% of Energiebetei ligungs-Gesellschaft. The ultimate owners are Bayernwerk AG, CONTIGAS and the State of Bavaria. Following the weakest link principle, control rights are allocated as follows: Bayernwerk AG 76% (51% plus 25%), CONTIGAS 26%, and State of Bavaria 1.7%.

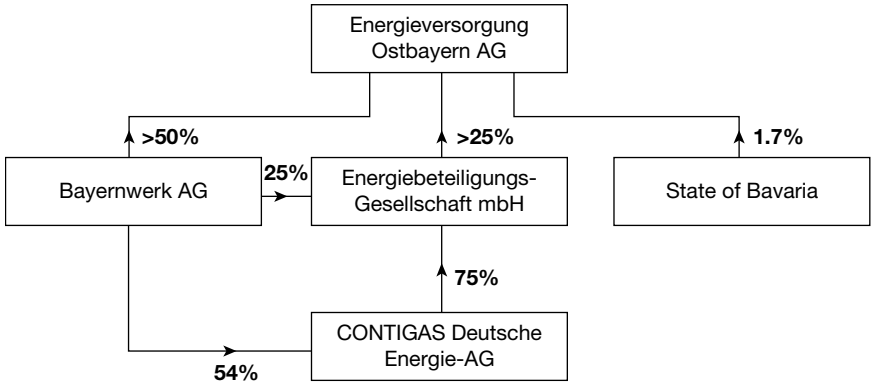


Figure 13B.1 Energieversorgung Ostbayern AG as an example of a complex pyramid with direct and indirect ownership, September 1986. Following our principle of defining control rights based on votes, the graph displays ownership as fractions of votes (which is not necessarily identical to the fractions of equity from which these votes emanate). Energieversorgung Ostbayern AG is owned by Bayernwerk AG (a nonfinancial firm) with more than 50% of the shares, Energiebeteiligungs-Gesellschaft mbH (a financial holding shell) with more than 25% of the shares, and the State of Bavaria with 1.7%. In addition, Bayernwerk owns 54% of CONTIGAS Deutsche Energie-AG, while CONTIGAS, in turn, owns 75% of Energiebeteiligungs-Gesellschaft. Bayernwerk also owns 35% of Energiebeteiligungs-Gesellschaft. Following the weakest link principle (La Porta et al., 1999a), control rights are allocated to the ultimate owners as follows: Bayernwerk AG 76% (51% plus 25%), CONTIGAS 26%, and State of Bavaria 1.7%. Data source: *Saling Aktienführer 1987,* Verlag Hoppenstedt, Darmstadt, 1986.

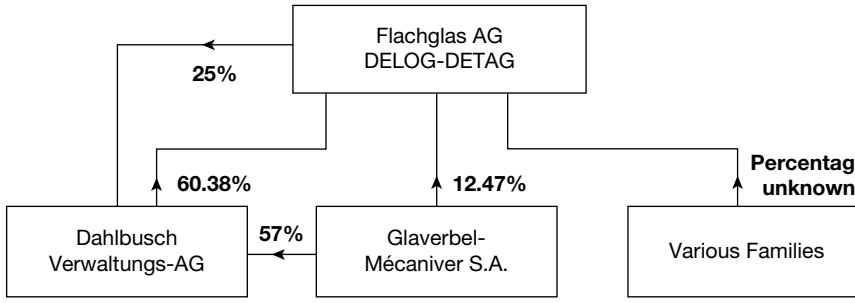


Figure 13B.2 Flachglas AG DELOG-DETAG as an example of a complex pyramid with direct and indirect ownership and cross-shareholding, September 1975. Following our principle of defining control rights based on votes, the graph displays ownership as fractions of votes (which is not necessarily identical to the fractions of equity from which these votes emanate). Flachglas AG DELOG-DETAG is owned by Dahlbusch Verwaltungs-AG, a domestic financial holding shell, with 60.38%, by Glaverbel-Mécaniver S.A., a Belgian nonfinancial firm, with 12.47%, and by various families with unknown percentages. Flachglas AG itself owns 25% of Dahlbusch (circularity). Another 57% of Dahlbusch Verwaltungs-AG is owned by Glaverbel-Mécaniver S.A. (i.e., Glaverbel-Mécaniver owns stakes in Flachglas directly and indirectly). Allocation of control rights according to the weakest link principle is as follows: Glaverbel-Mécaniver S.A. is allocated 69.47% (57% plus 12.47%) and the firm itself (i.e., Flachglas AG) is allocated 25%. Data source: *Saling Aktienführer 1976,* Verlag Hoppenstedt, Darmstadt, 1975.

The second example, shown in Figure 13B.2, shows a pyramid with indirect ownership, direct ownership and circular ownership. (Again, the graph displays ownership as fractions of votes, which is not necessarily identical to ownership of equity.) In September 1975, Flachglas AG DELOGDETAG was owned by Dahlbusch Verwaltungs-AG, a domestic financial holding shell, with 60.38%, by Glaverbel-Mécaniver S.A., a Belgian nonfinancial firm, with 12.47%, and by various families with unknown percentages. Flachglas AG itself owns 25% of Dahlbusch (circularity). About another 57% of Dahlbusch Verwaltungs-AG is owned by Glaverbel-Mécaniver S.A. (i.e., Glaverbel-Mécaniver owns stakes in Flachglas directly and indirectly). (We do not know the percentages of the families simply because they are not reported by Hoppenstedt. We use the term "about 57%" because Hoppenstedt uses it.) Allocation of control rights according to the weakest link principle is as follows: Glaverbel-Mécaniver S.A. is allocated 69.47% (57% plus 12.47%) and the firm itself (i.e., Flachglas AG) is allocated 25%.

REFERENCES

Admati, A., Pfleiderer, P., Zechner, J., 1994. Large shareholder activism, risk sharing, and financial market equilibrium. Journal of Political Economy 102, 1097–1130.

Allen, F., Gale, D., 1997. Financial markets, intermediaries, and intertemporal smoothing. Journal of Political Economy 105, 523–46.

Banzhaf, J.G., 1965. Weighted voting doesn't work: a mathematical analysis. Rutgers Law Review 19, 317–43.

Banzhaf, J.G., 1968. One man, 3.312 votes: a mathematical analysis of the Electoral College. Villanova Law Review 13, 304–22.

Barclay, M., Holderness, C., 1991. Negotiated block trades and corporate control. Journal of Finance 46, 861–78.

Bethel, J., Liebeskind, J.P., Opler, T., 1998. Block share purchases and corporate performance. Journal of Finance 53, 605–34.

Bhide, A., 1993. Hidden cost of stock market liquidity. Journal of Financial Economics 34, 31–51.

Böhm, J., 1992. Der Einfluß der Banken auf Großunternehmen. Steuer-und Wirtschaftsverlag, Hamburg.

Cable, J.R., 1985. Capital market information and industrial performance: the role of West German banks. Economic Journal 95, 118–32.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74, 829–36.

Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association 83, 596–610.

Cleveland, W.S., Devlin, S.J., Grosse, E., 1988. Regression by local fitting: methods, properties, and computational algorithms. Journal of Econometrics 37, 87–114.

Coenenberg, A.G., 1974. Jahresabschluß und Jahresabschlußanalyse. Verlag Moderne Industrie, München.

Coenenberg, A.G., 1993. Jahresabschluß und Jahresabschlußanalyse, 14th edition. Verlag Moderne Industrie, Landsberg am Lech.

Demsetz, H., Lehn, K., 1985. The structure of corporate ownership: causes and consequences. Journal of Political Economy 93, 1155–77.

Dow, J., Gorton, G., 1997. Stock market efficiency and economic efficiency: is there a connection? Journal of Finance 52, 1087–129.

Edwards, J., Fischer, K., 1994. Banks, Finance and Investment in Germany. Cambridge University Press, Cambridge.

Eicker, 1967. Limit theorems for regressions with unequal and dependent errors. In: Le Cam, L., Neyman, J. (Eds.), Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, 59–82.

Elsas, R., Krahnen, J.P., 1998. Is relationship lending special? Evidence from credit-file data in placeGermany. Journal of Banking and Finance 22, 1283–316.

Emmons, W., Schmid, F., 1998. Universal banking, control rights, and corporate finance in Germany. Federal Reserve Bank of St. Louis Review 80, 19–42.

Franks, J., Mayer, C., 2000. Ownership and control of German corporations. Unpublished working paper. London Business School and University of Oxford.

Gorton, G., Rosen, R., 1995. Corporate control, portfolio choice, and the decline of banking. Journal of Finance 50, 1377–420.

Gorton, G., Schmid, F., 1998. Corporate finance, control rights, and firm performance: a study of German codetermination. Unpublished working paper. Wharton School, Pennsylvania.

Gottschalk, A., 1988. Der Stimmrechtseinfluß der Banken in den Aktionärsversammlungen der Großunternehmen, WSI-Mitteilungen 41, 294–304.
Grossman, S., Hart, O., 1980. Takeover bids, the free-rider problem, and the theory of the corporation. Bell Journal of Economics 11, 42–64.
Grossman, S., Hart, O., 1988. One share-one vote and the market for corporate control. Journal of Financial Economics 20, 175–202.
Grundfest, J., 1990. Subordination of American capital. Journal of Financial Economics 27, 89–114.
Halvorsen, R., Palmquist, R., 1980. The interpretation of dummy variables in semilogarithmic equations. American Economic Review 70, 474–75.
Harris, M., Raviv, A., 1988. Corporate governance: voting rights and majority rules. Journal of Financial Economics 20, 203–35.
Harris, T., Lang, M., Möller, H.P., 1994. The value relevance of German accounting measures: an empirical analysis. Journal of Accounting Research 32, 187–209.
Herrhausen, A., 1987. Kontroverseüber die Macht der Banken. Verbraucherpolitische Hefte 1987(5), 99–109.
Hilferding, R., 1910. Das Finanzkapital. Dietz, Berlin.
Holderness, C., Sheehan, D., 1988. The role of majority shareholders in publicly held corporations: an exploratory analysis. Journal of Financial Economics 20, 317–46.
Jensen, M., Meckling, W., 1976. Theory of the firm: managerial behavior, agency costs, and ownership structure. Journal of Financial Economics 3, 305–60.
Kahn, C., Winton, A., 1998. Ownership structure, speculation, and shareholder intervention. Journal of Finance 53, 99–129.
Körber, U., 1989. Die Stimmrechtsvertretung durch Kreditinstitute. Duncker & Humblot, Berlin.
Krümmel, H.J., 1980. German universal banking scrutinized: some remarks concerning the Gessler report. Journal of Banking and Finance 4, 33–55.
La Porta, R., Lopez-de-Silanes, F., Shleifer, A., 1999. Corporate ownership around the world. Journal of Finance 54, 471–517.
La Porta, R., Lopez-de-Silanes, F., Shleifer, A., Vishny, R., 1999. Investor protection and corporate valuation. Harvard Economic Institute, Discussion Paper Number 1883.
Leech, D., 1988. The relationship between shareholding concentration and shareholding voting power in British companies: a study of the application of power indices for simple games. Management Science 34, 509–26.
Leech, D., Leahy, J., 1991. Ownership structure, control type classifications and the performance of large British companies. Economic Journal 101, 1418–37.
Mallows, C.L., 1973. Some comments on C_p. Technometrics 15, 661–75.
Maug, E., 1998. Large shareholders as monitors: is there a trade-off between liquidity and control? Journal of Finance 53, 65–98.
McConnell, J., Servaes, H., 1990. Additional evidence on equity ownership and corporate value. Journal of Financial Economics 27, 595–612.
Mikkelson, W., Ruback, R., 1985. An empirical analysis of the interfirm equity investment process. Journal of Financial Economics 14, 523–53.
Monopolkommission, 1977. Hauptgutachten I (1973/75), Mehr Wettbewerb ist möglich. Nomos, Baden-Baden.
Monopolkommission, 1978. Hauptgutachten II (1976/77), Fortschreitende Konzentrationbei Großunternehmen. Nomos, Baden-Baden.

Monopolkommission, 1980. Hauptgutachten III (1978/79), Fusionskontrolle bleibt vorrangig Nomos, Baden-Baden.
Morck, R., Shleifer, A., Vishny, R., 1988. Management ownership and market valuation: an empirical analysis. Journal of Financial Economics 20, 293–315.
Müller, H.G., 1987. Weighted local regression and kernel methods for nonparametric curve fitting. Journal of the American Statistical Association 82, 231–38.
Ordelheide, D., Pfaff, D., 1994. Germany, European Financial Reporting Series Routledge, New York and London.
Perfect, S., Wiles, K., 1994. Alternative construction of Tobin's Q: an empirical comparison. Journal of Empirical Finance 1, 313–41.
Porter, M., 1992. Capital choices: changing the way America invests in industry. Journal of Applied Corporate Finance 5(2), 4–16.
Shapley, L.S., Shubik, M., 1954. A method for evaluating the distribution of power in a committee system. American Political Science Review 48, 787–92.
Shleifer, A., Vishny, R., 1986. Large shareholders and corporate control. Journal of Political Economy 94, 461–88.
Speckman, P., 1988. Kernel smoothing in partial linear models. Journal of the Royal Statistical Association, Series B, 50, 413–36.
Spiro, 1958. The Politics of Codetermination. Harvard University Press, Massachusetts.
Studienkommission, 1979. Bericht der StudienkommissionGrundsatzfragen der Kreditwirtschaft. Schriftenreihe des Bundesministeriums für Finanzen, Heft 28.Stollfuß, Bonn.
Stute, W., 1984. Asymptotic normality of nearest neighbor regression function estimates. Annals of Statistics 12, 917–26.
United Nations, 1990. International standard industrial classification of all economic activities. Statistical Papers, Series M, No. 4, Rev. 3, New York.
Wenger, E., Kaserer, C., 1998. The German system of corporate governance—a model that should not be imitated. In: Black, S.W., Moersch, M. (Eds.), Competition and Convergence in Financial Markets. Amsterdam, Elsevier, 41–78.
Whang, Y.J., Andrews, D., 1993. Tests of specification for parametric and semiparametric models. Journal of Econometrics 57, 277–318.
White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–38.
Wiedemann, H., 1980. Codetermination by workers in German enterprises. The American Journal of Comparative Law 28, 79–82.
Zeckhauser, R., Pound, J., 1990. Are large shareholders effective monitors? An investigation of share ownership and corporate performance. In Hubbard, G. (Ed.), Asymmetric Information, Corporate Finance, and Investment. University of Chicago Press, Chicago, 149–180.

14

Bank Credit Cycles

GARY B. GORTON AND PING HE*

14.1. INTRODUCTION

The essence of banking is the determination as to whether a potential borrower is credit-worthy, that is, whether the potential borrower meets the bank's credit standards. When each bank makes this determination, it does so in competition with other banks, each with its own proprietary lending standards. In this paper we analyze this bank competition, presenting a repeated game of bank lending, in the style of Green and Porter (1984), in which banks can change their lending standards. In the theoretical model, we show that the bank competition for borrowers leads to periodic credit crunches, swings between high and low credit allocations. The reason is that bank lending standards vary through time due to strategic interaction between competing banks. Credit cycles can occur without any change in the macroeconomic environment.

We then go on to empirically investigate this lending standard model, providing empirical evidence that bank credit cycles are an important autonomous part of business cycle dynamics. Empirical tests take advantage of the unique information environment in U.S. banking, where detailed information about rival banks is collected and released periodically by the bank regulators. Thus, the information that is the basis for banks' beliefs about rival banks' lending standards is observable to the econometrician. This allows for a novel approach

* We thank Yacine Ait-Sahalia, Bernard Salanié, Kent Daniel, Steve Davis, Xavier Gabaix, Armando Gomes, Charles Kahn, Anil Kashyap, Richard Kihlstrom, Narayana Kocherlakota, Rob McMillan, George Mailath, Stewart Mayhew, Ben Polak, Eric Rosengren, Geert Rouwenhorst, José Scheinkman, Hyun Shin, Nick Souleles, Jeremy Stein, two anonymous referees, and seminar participants at Princeton, Yale, Duke, Chicago, Tsinghua, the New York Federal Reserve Bank, the U.S. Securities and Exchange Commission, Moody's Investors Services and the NBER Conference on Capital Markets and the Economy for their comments and suggestions.

to testing the repeated game. We propose direct measures of the information that the theory suggests are relevant for banks' beliefs. We use these measures as proxies for the beliefs themselves and show how these proxies drive the credit cycle.

Bank lending is clearly an important topic. Changes in bank credit allocation, sometimes called "credit crunches," appear to be an important part of macroeconomic dynamics. Bank lending is procyclical.[1] Rather than change the price of loans, the interest rate, banks sometimes ration credit.[2] A dramatic example in the U.S. is the period shortly after the Basel Accord was agreed in 1988, during which time the share of U.S. total bank assets composed of commercial and industrial loans fell from about 22.5 percent in 1989 to less than 16 percent in 1994. At the same time, the share of assets invested in government securities increased from just over 15 percent to almost 25 percent.[3] More generally, it has been noted that banks vary their lending standards or credit standards.

Bank "lending standards" or "credit standards" are the criteria by which banks determine and rank loan applicants' risks of loss due to default, and according to which a bank then makes its lending decisions. While not observable, there is a variety of evidence showing that while lending rates are sticky, banks do, in fact, change their lending standards.[4] The most direct evidence comes from the Federal Reserve System's Senior Loan Officer Opinion Survey on Bank Lending Practices.[5] Banks are asked whether their "credit standards" for approving loans (excluding merger and acquisition-related loans) have "tightened considerably, tightened somewhat, remained basically unchanged, eased somewhat, or eased

1. See Lown, Morgan and Rohatgi (2000), Jordan, Peek, and Rosengren (2002), and Lown and Morgan (2002).

2. Bank loan rates are sticky. Berger and Udell (1992) regress loan rate premiums against open market rates and control variables and find evidence of "stickiness." (Also, see Berger and Udell (1992) for references to the prior literature.) With respect to credit card rates, in particular, Ausubel (1991) has also argued that they are "exceptionally sticky relative to the cost of funds" (p. 50).

3. See Keeton (1994) and Furfine (2001). This episode is the focus of the empirical literature on credit crunches. See Bernanke and Lown (1991), Hall (1993), Berger and Udell (1994), Haubrich and Wachtel (1993), Hancock and Wilcox (1994), Brinkman and Horvitz (1995), Peek and Rosengren (1995), and Beatty and Gron (2001). Gorton and Winton (2002) provide a brief survey of the credit crunch literature.

4. In the absence of detailed information about banks' internal workings, it is not exactly clear what is meant by the term "lending standards." It can refer to all the elements that go into making a credit decision, including credit scoring models, the lending culture, the number of loan officers and their seniority and experience, the banks' hierarchy of decision-making, and so on.

5. The survey is conducted quarterly and covers major banks from all parts of the U.S., accounting for between 60 and 70 percent of commercial and industrial loans in the U.S. The Federal Reserve System's "Senior Loan Officer Opinion Survey on Bank Lending Practices" was initiated in 1964,

considerably." Lown and Morgan (2005) examine this survey evidence and note that, except for 1982, every recession was preceded by a sharp spike in the net percentage of banks reporting a tightening of lending standards. Other evidence that bank lending standards change is econometric. Asea and Blomberg (1998) examined a large panel data set of bank loan terms over the period 1977 to 1993 and "demonstrate that banks change their lending standards—from tightness to laxity—systematically over the cycle" (p. 89). They concluded that cycles in bank lending standards are important in explaining aggregate economic activity.

Also in a macroeconomic context, changes in the Fed Lending Standards Index (the net percentage of respondents reporting tightening) Granger-causes changes in output, loans, and the federal funds rate, but the macroeconomic variables are not successful in explaining variation in the Lending Standards Index.[6] The Lending Standards Index is exogenous with respect to the other variables in the Vector Autoregression system. See Lown and Morgan (2005, 2002) and Lown, Morgan and Rohatgi (2000).[7] The analysis in this paper is aimed at explaining the forces that cause lending standards to change and, in particular, to explain how this can happen independently of macroeconomic variables.

To investigate bank lending standards we construct a model of bank lending that is predicated on the special features of banks, namely, that banks produce private information about potential borrowers when they determine whether borrowers meet their lending standards. Broecker (1990) emphasizes that this information asymmetry means that banks compete with each other in a special way. When competing with each other to lend, banks produce information about potential borrowers in an environment where they do not know how much information is being produced by rival bank lenders.[8] We study a repeated model of bank competition, a la Green and Porter (1984), in which banks collude to set high loan rates (hence loan rates are sticky), and they implicitly agree not to

but results were only made public starting in 1967. Between 1984:1 and 1990:1 the question concerning lending standards was dropped. See Schreft and Owens (1991). Current survey results are available at <http://www.federalreserve.gov/boarddocs/SnLoanSurvey/>.

6. Lown and Morgan (2002, 2005) use the survey results to create an index: the number of loan officers reporting tightening standards less the number of reporting loan officers reporting easing standards divided by the total number reporting.

7. They also find that changes in bank lending standards matter much more for the volume of bank loans and aggregate output than do commercial loan rates, consistent with the finding that loan rates do not move as much as would be dictated by market rates.

8. In Broecker's (1990) model, banks use noisy, independent, credit worthiness tests to assess the riskiness of potential borrowers. Because the tests are imperfect, banks may mistakenly grant credit to high-risk borrowers whom they would otherwise reject. As the number of banks increases, the likelihood that an applicant will pass the test of at least one bank rises. Banks face an inherent

(over-) invest in costly information production about prospective borrowers.[9] A bank can strategically produce more information than its rivals and then select the better borrowers, leaving unknowing rivals with adversely selected loan portfolios. Unlike standard models of imperfect competition, following Green and Porter (1984), there are no price wars among banks since banks do not change their loan rates. However, as in Green and Porter (1984), intertemporal incentives to maintain the collusive arrangement requires periods of "punishment." Here these correspond to credit crunches. In a credit crunch all banks increase their costly information production intensity, that is, they raise their "lending standards," and stop making loans to some borrowers who previously received loans. These swings in credit availability are caused by banks' changing beliefs, based on public information about rivals, about the viability of the collusive arrangement.

Repeated games are difficult to test and that is the case here.[10] There are many equilibria, depending on agents' beliefs. Agents' beliefs about other agents' beliefs depend on current information and the history of the game. We empirically determine the equilibrium, i.e., "test" the model, by parameterizing the public information that is the basis for banks' beliefs about rivals' strategies, and using such measures as proxies for beliefs. The empirical behavior of U.S. bank credit card lending, commercial and industrial lending, and bank profitability, are consistent with the model. Bank credit cycles are a systematic risk. We find that, consistent with this, our belief proxy, called the Performance Difference Index (*PDI*), as explained later, is a priced factor in an asset pricing model of bank stock returns. Most importantly, the *PDI* is a priced factor for non-financial firms as well and increasingly so as firm size declines.

We show theoretically that to detect deviations by rival banks, each bank looks at two pieces of public information: the number of loans made in the period by each rival and the default performance of each rivals' loan portfolio. This is an implication of banks competing using information production intensity (lending standards). The relative performance of other banks is the public information

winner's curse problem in this setting. In Broecker's model banks do not behave strategically in a dynamic way.

9. Strategic interaction between banks seems natural because banking is highly concentrated. Entry into banking is restricted by governments. In developed economies the share of the largest five banks in total bank deposits ranges from a high of 81.7% in Holland to a low of 26.3% in the United States. See the Group of Ten (2001). In less developed economies, bank concentration is typically much higher (see Beck, Demirguc-Kunt, and Levine (2003)).

10. Empirically testing models of repeated strategic interaction of firms has focused on price wars. See Reiss and Wolak (2003) and Bresnahan (1989) for surveys of the literature. However, our model predicts that there are "information production wars." Since information production is unobservable, we cannot follow the usual empirical strategy. We propose a new method for empirically investigating such models.

relevant for each bank's decisions about the choice of the level of information production. Intuitively, excessive information production by a bank will not change the overall loan performance on average, but will change the distribution of loan defaults across banks. Moreover, the use of relative bank performance empirically distinguishes our theory from a general learning story, which would predict past bank performance matters for bank credit decisions (an alternative hypothesis which we test).

Broadly, the empirical analysis is in three parts. First, we examine a narrow category of loans, U.S. credit card lending, where there are a small number of banks that appear to dominate the market. Even with a small number of banks it is not obvious which banks are rivals, so we first analyze this lending market by examining banks pairwise. If the *PDI* increases, banks should reduce their lending and increase their information production resulting in fewer loan losses in the next quarter. We also examine big credit card lender banks' profitability, using stock returns.

Second, we turn to the macro economy by looking at commercial and industrial loans. We analyze a number of macroeconomic time series, including the Lending Standard Survey Index. We form an aggregate bank Performance Difference Index based on the absolute value of the differences on all commercial and industrial loans of the largest 100 banks. If beliefs are, in fact, based on this information, then we should be able to explain (in the sense of Granger causality) the time series behavior of the Lending Standard Survey responses (the percentage of banks reporting "tightening" their standards).

Finally, if credit crunches are endogenous, and a systematic risk, then they should be a priced factor in an asset pricing model of stock returns. Therefore, our final test is to ask whether a mimicking portfolio for our parameterization of banks' relevant histories is a priced risk factor in a CAPM or Fama-French asset pricing setting. We look at banks and nonfinancial firms by size, as credit crunches have larger effects on smaller firms. We find the evidence to be consistent with the theory.

Two related theoretical models are provided by Dell'Ariccia and Marquez (2004) and Ruckes (2003). These papers show a link between lending standards and information asymmetry among banks, driven by exogenous changes in the macroeconomy. As distinct from these models, the fluctuation of banks' lending behavior in our paper is purely driven by the strategic interactions between banks instead of an exogenously changing economic environment.

In terms of empirical work, Rajan (1994) is related. He argues that fluctuations in credit availability by banks are driven by bank managers' concerns for their reputations (due to bank managers having short horizons), and that consequently bank managers are influenced by the credit policies of other banks. Managers' reputations suffer if they fail to expand credit while other banks are

doing so, implying that expansions lead to significant increases in losses on loans subsequently.[11] We test Rajan's idea in the empirical section.

Also related to our work, though more distantly, is some research in Monetary Theory, in particular on the "bank lending channel."[12] The "bank lending channel" posits that disruptions in the supply of bank loans can be caused by monetary policy, resulting in credit crunches (see Bernanke and Blinder (1988)). If bank funding is interest rate sensitive, then perhaps changes in banks' cost of funds results in variation in the amount of credit that banks supply. The bank lending channel is controversial because, as some have argued, banks have access to non-deposit sources of funds. See Ashcraft (2003) for evidence against the bank lending channel. We do not investigate the effects of monetary policy here, though this is a topic for future research. We provide the micro foundations for how bank competition can cause credit crunches independent of monetary policy, but this is not mutually exclusive from the bank lending channel. However, like the bank lending channel literature, we assume that there are no perfect substitutes for bank loans, so that if borrowers are cut off from bank credit they cannot find alternative financing at the same price, especially small firms. Large firms usually have access to capital markets.

We proceed in Section 14.2 to first describe the stage game for bank lending competition, and we study the existence of stage Nash equilibrium and the model's implications for lending standards, and the stage game is followed by repeated competition. In Section 14.3, we carry out empirical tests in the credit card loan market, a market dominated by a small number of banks. In Section 14.4 we extend the empirical analysis to commercial and industrial loans, the most important category of loans. We test whether our model can explain credit crunches. Section 14.5 undertakes a different type of test. We ask whether the risk caused by bank strategic behavior is priced in an asset pricing context. Finally, Section 14.6 concludes the paper.

14.2. THE LENDING MARKET GAME

We first set forth the lending market stage game. To simplify our discussion, suppose that there are two banks in the market competing to lend, as follows. There

11. However, as pointed out by Weinberg (1995), the data on the growth rate of total loans and loan charge-offs in the United States from 1950 to 1992 do not show the pattern of increases in the amount of lending being followed by increases in loan losses.

12. The credit channel of monetary policy transmission has focused on the two ways that central bank action can affect real economic activity by increasing the "external finance premium" (see Bernanke and Gertler (1995) for a review). One of these is the "balance sheet channel," which is concerned with effects of monetary policy on firms' credit worthiness. Increases in interest rates, for example, may reduce the value of the collateral that firms borrow against. The other is the "bank lending channel," which is more relevant for our work.

are N potential borrowers in the credit market. Each of the potential borrowers is one of two types, good or bad. Good types' projects succeed with probability p_g, and bad types' projects succeed with probability p_b, where $p_g > p_b \geq 0$. Potential borrowers, sometimes also referred to below as "applicants," do not know their own type. At the beginning of the period potential borrowers apply simultaneously to each bank for a loan. There is no application fee. The probability of an applicant being a bad type is λ, which is common knowledge.[13] Each applicant can accept at most one loan offer, and if a loan is granted, the borrower invests in a one period project which will yield a return of $X < \infty$ if the project succeeds and returns 0 otherwise. A borrower whose project succeeds will use the return X to repay the loan, i.e., a borrower's realized cash flow is verifiable.

Banks are risk-neutral. They can raise funds at some interest rate, assumed to be zero. After receiving the loan applications, a bank can use a costly technology to produce information about the applicant's type. The credit worthiness testing results in determining the type of an applicant, but there is a per applicant cost of $c > 0$. Banks can test any proportion of their applicants. Let n_i denote the number of applicants that are tested by bank i. We say that the more applicants that a bank tests, using the costly information production technology, the higher are its credit or lending standards.[14] If a bank switches from not using the credit worthiness test to using it, or tests more applicants, we say that the bank has "raised" its lending or credit standards. We assume that neither bank observes the other bank's credit standards, i.e., each bank is unaware of how many applicants the other bank tests. Results of the tests are the private information of the testing bank.

Since the bank borrowing rate is zero, when a bank charges F (to be repaid at the end of the period) for one unit of loan, the bank's expected return from lending to an applicant will be $\lambda p_b F + (1-\lambda)p_g F - 1$ in the case of no credit worthiness testing. We assume:

Assumption 1: $p_g X > 1$, $p_b X < 1$, and $\lambda p_b X + (1-\lambda)p_g X > 1$.

Assumption 1 means that there exists some interest rate, X, that allows a bank to earn positive profits from lending to a good type project ex ante, but there does not exist an interest rate at which a bank can make positive profits from lending to a bad type project ex ante. (Given the loan size being normalized to

13. We will hold λ fixed throughout the analysis, but this is to clarify the mechanism that is our focus. It is natural to think of λ as being time-varying, representing other business cycle shocks outside the model, and we could easily incorporate this. But it would obscure the cyclical effects that are purely due to bank competition.

14. Imagine that banks always produce some minimal amount of information about loan applicants. We ignore this base amount of information, however, and focus only on the situation where banks choose to produce more information than this base level. So, we interpret the credit worthiness test as the additional information produced, beyond the normal information production.

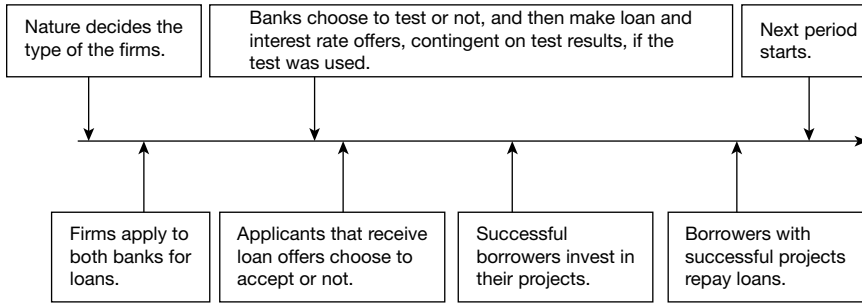


Figure 14.1 The Timing of the Stage Game

1, the face value of the loan F uniquely determines the interest rate, and later on we refer to F as the "loan interest rate.") It is also possible for banks to profit from lending to both types of applicants without discriminating between the types.

Each bank first chooses some (possibly none, possibly all) applicants to test, then, depending on the test results, decides whether to make a loan offer for each applicant, and if yes, at what interest rate. We formally define the stage strategy of each bank in the Appendix A. We assume that banks do not observe each other's interest rates or the identities of applicants offered loans. At the end of the period only final loan portfolio sizes and loan outcomes (i.e., default or not) are publicly observable. Banks cannot communicate with each other. Figure 14.1 shows the timing of the stage game.

14.2.1. Stage Nash Equilibrium

We now turn to study Nash equilibrium, and the conditions for the existence of Nash equilibrium, in the lending market stage game. We show that in the stage game, banks have no incentive to conduct the credit worthiness tests, and we provide a condition under which the only Nash equilibrium that exists is one in which neither bank conducts credit worthiness testing and both banks earn zero profits.

First we will study the Nash equilibrium in which no bank conducts credit worthiness testing. The following assumption guarantees the existence of such equilibria.

Assumption 2: $c \geq \frac{\lambda(1-\lambda)(p_g-p_b)}{\lambda p_b+(1-\lambda)p_g}$.

Assumption 2 also implies that the optimal payoffs for the banks are reached when no credit worthiness testing are conducted (as we will show later).

PROPOSITION 1 *Under Assumption 2, there exists a symmetric Nash equilibrium in which no bank conducts credit worthiness testing and both banks earn zero profits.*

The proof is in Appendix B.

Proposition 1 says that if the cost of testing each loan applicant is sufficiently high, then there exists a Nash equilibrium in which no bank conducts credit worthiness testing and neither bank earns positive profits.

Now consider the case where both banks test at least some applicants.

PROPOSITION 2: *There is no symmetric Nash equilibrium in which both banks test at least some of the applicants.*

The proof is in Appendix C.[15]

Intuitively, after the banks test some of the applicants, they will compete with each other for the good type applicants, which will drive the post-test profit to zero. However, since there is a test cost, ex-ante the banks' profits will be negative.

Our conclusion with regard to the stage game in the lending market is that, without mixed strategies, the only Nash equilibrium that exists is the equilibrium in which neither bank conducts credit worthiness testing, and both banks earn zero profits.

It is straightforward to characterize the optimal payoffs that the two banks receive in the stage game. If a bank does not conduct credit worthiness testing on an individual applicant and charges F, then the expected payoff from a loan to that individual applicant is $\pi = \lambda p_b F + (1-\lambda)p_g F - 1$, which is maximized at $F = X$. If a bank conducts credit worthiness testing on an individual applicant and charges F, then the expected payoff from a loan to that individual applicant is $\pi' = (1-\lambda)p_g F - 1 - c$, which also is maximized at $F = X$. It is easy to check that $\pi' < \pi$ with $F = X$ under Assumption 2.

14.2.2. Repeated Competition

We formalize the repeated game in Appendix D. In the stage game, we have already shown that banks earn zero profits without testing, and the optimal payoffs for banks are reached when there is no costly credit worthiness test being used. Setting a (collusive) loan interest rate of $F = X$ would be the most profitable case for both banks. Ideally, in repeated competition banks will try to collude to charge $F = X$ without conducting credit worthiness testing. When the banks collude by offering a profitable interest rate to the applicants without testing, there is an incentive for each bank to undercut the interest rate in order to get more applicants. In order to generate intertemporal incentives to support

15. Banks could play more general mixed strategies. For example, banks could mix between testing n_1 applicants and testing n_2 applicants. We do not delve into these strategies.

the collusion on a high interest rate, banks need to punish each other to prevent deviation in undercutting interest rates, which can be monitored by looking at the loan portfolio size of each bank. However, a high interest rate generates incentives for banks to conduct credit worthiness testing and get higher quality applicants while manipulating the loan portfolio size. To see this, let us look at the following example.

By undercutting the interest rate offered to an applicant without credit worthiness testing, the expected payoff from this loan to the bank is: $\pi = \lambda p_b F + (1-\lambda)p_g F - 1$. Alternatively, the bank can test the applicant, undercut the interest rate if it is a good type, and undercut the interest rate to another untested applicant if the tested one turns out to be a bad type (this way the bank always gets one applicant for sure); the expected payoff to the bank is $\pi'' = \lambda[\lambda p_b F + (1-\lambda_g F) - 1] + (1-\lambda)(p_g F - 1) - c$. The difference between π'' and π is $\lambda(1-\lambda)(p_g - p_b)F - c$, which is increasing with F. When there are multiple applicants, while benefiting from finding a good type applicant through a credit worthiness test, a bank will switch to an untested applicant if the tested one turns out to be of bad type, and this substantially improves the net gain from a credit worthiness test. Therefore, when F is high enough, banks will have an incentive to produce information while manipulating the loan portfolio size through interest rates. To proceed, we make the following assumption:

Assumption 3: $c \leq \lambda(1-\lambda)(p_g - p_b)X$.

This assumption guarantees that when banks collude at the highest possible interest rate, X, they have incentive to over-produce information and undercut interesting rates.

Aside from seeing how the repeated game works, the main point is the demonstration that because banks have two actions that they can use to compete (i.e., changing lending rates and increasing information production), banks' beliefs must be based on the history of banks' portfolio sizes as well as banks' loan default performances.

At a profitable interest rate, if a bank makes more loans than its rival, then the continuation value of this bank should be lower, to eliminate the incentive of the banks to deviate by undercutting interest rates to get more loans. However, when there is credit worthiness testing, it may not be true that making more loans is always better. A bank can deviate by testing, "raising credit standards," resulting in the other bank lending to the bad type applicants rejected by the first bank. This is the strategic use of the winner's curse by one bank against its rival. Due to that possibility, it is easy to imagine (and we can formally show) that loan performance (number of defaults in each bank portfolio) will also affect the continuation value. When the banks want to avoid costly credit worthiness testing on the equilibrium path, then it is not possible for the two banks to collude on a high loan interest rate in equilibrium without looking

at each other's loan performances. The possibility of deviating by using credit worthiness testing while manipulating the loan size, and the resulting winner's curse effect, makes both banks' strategies sensitive to each others' past loan performances, even though there is an i.i.d. distribution of borrower types over time.

To demonstrate that monitoring through loan size only is not sufficient to detect a deviation, let us first look at an example with two loan applicants, where each bank makes a loan offer to both loan applicants at interest rate $F_\alpha > F^* = \frac{1}{\lambda p_b + (1-\lambda)p_g}$ without a credit worthiness test. Consider a deviation to a strategy in which a bank tests one applicant. If the tested applicant is a bad type the bank rejects it and, without testing the other applicant, undercuts the interest rate to F_α^- for the loan to the other applicant. If the tested applicant is a good type then the bank offers a loan to the applicant at F_α^- and raises the interest rate to F_α^+ for the loan to (or rejects) the other untested applicant. In this way the expected loan portfolio size for both banks will remain the same while the distribution of the loan portfolio size changes a little. It is easy to check that the improvement in the stage profit for the deviating bank is $\Delta E[\pi] = -c + \lambda(1-\lambda)(p_g - p_b)F_\alpha$, and $\Delta E[\pi] > 0$ as long as F_α is close enough to X, by Assumption 3.

In our example with two loan applicants, if one bank deviates in the way we described above, then the loan allocation is $(1, 1)$ with probability 1, while without a deviation, the loan allocation is $(2, 0)$ with probability 0.25, $(1, 1)$ with probability 0.5, and $(0, 2)$ with probability 0.25. Let $u_i(n_1, n_2)$ denote the payoff to bank i when the loan allocation is (D_1, D_2), and we know by Lemma 5 in Appendix E that, in a Symmetric Perfect Public Equilibrium:

$$u_1(0,2) - u_1(1,1) = u_1(1,1) - u_1(2,0),$$

which implies:

$$0.25u_1(0,2) + 0.5u_1(1,1) + 0.25u_1(2,0) = u_1(1,1).$$

Thus with the deviation, the expected continuation payoff remains unchanged. We can show that this result holds with more than two applicants for any Symmetric Perfect Public Equilibrium, as defined in the Appendix; we omit the proof here for brevity.

Therefore, in order to detect banks' deviations through over-production of information, banks' strategies need to depend on the public histories of banks' loan portfolio performances and portfolio sizes. However, the theory does not provide details on how the public histories are linked to banks' beliefs and strategies. To help understand this issue for later empirical tests, let us again consider a simple example with $N = 2$ applicants. Suppose Bank 1 deviates from the equilibrium strategy s (test no applicants, and offer some high interest rate F_α

to both of them) to strategy s' as follows: test one applicant; if he is good, offer a loan at rate F_α^-, and reject the other applicant; if the applicant is bad, reject it, and offer a loan to the other applicant at loan rate F_α^-. In this way, the expected loan portfolio size is not changed, but loan performance will be improved; there is less likely to be a default. Given the loan distribution $(D_1 = 1,\ D_2 = 1)$, from Bank 2's point of view, without deviation by Bank 1, the probability of Bank 2 having a loan default is:

$$q = \lambda(1 - p_b) + (1 - \lambda)(1 - p_g).$$

With Bank 1 deviating to strategy s', Bank 2's default probability becomes:

$$q' = \lambda(1 - p_b) + (1 - \lambda)[\lambda(1 - p_b) + (1 - \lambda)(1 - p_g)].$$

The likelihood of default is higher by:

$$\Delta q = q' - q = \lambda(1 - \lambda)(p_g - p_b) < 0.$$

To detect a deviation, however, banks should compare their results. That is, they should check their loan performance difference. Given the loan distribution $(D_1 = 1, D_2 = 1)$, without deviation by Bank 1, the probability of Bank 2 having a worse performance than Bank 1 is:

$$q_r = [\lambda(1 - p_b) + (1 - \lambda)(1 - p_g)][\lambda p_b + (1 - \lambda)p_g] < q.$$

With Bank 1 deviating to strategy s', this probability becomes:

$$q_r' = \lambda(1 - p_b)[\lambda p_b + (1 - \lambda)p_g] + (1 - \lambda)[\lambda(1 - p_b) + (1 - \lambda)(1 - p_g)]p_g.$$

We have:

$$\Delta q_r = q_r' - q_r = \lambda(1 - \lambda)(p_g - p_b) = \Delta q.$$

Therefore, compared with punishing each other after a bad performance, doing that after a relatively bad performance incurs a smaller probability of a mistaken punishment $\left(q_r < q\right)$, while it generates the same incentive to not to deviate $\left(\Delta q_r = \Delta q\right)$. The measure of the "performance difference" excludes the case where both banks perform poorly, and excluding this case is empirically important because it can result from aggregate shocks, which we do not model, and which does not differentiate our story from other alternative stories such as learning effect.

Before we start our empirical section, let us briefly discuss the link between information production and credit crunches. When each bank tests a subset of the applicant pool, the winner's curse effect may lead the banks to reject all those non-tested applicants. To see this, assume the banks randomly pick $n < N$ applicants for testing, and offers loans to those that pass the test. To simplify

the argument, assume that the interest rates offered to non-tested applicants are higher than the one offered to applicants that passed the test. For the non-tested applicants, it is possible that there does not exist a profitable interest rate due to the winner's curse. If a bank offers loans to non-tested applicants, then given an offer is accepted by an applicant, the probability of this non-tested applicant being a bad type is:

$$\theta = \Pr(\text{bad type} \mid \text{not tested}) = \frac{\frac{n}{N}\lambda + (1 - \frac{n}{N})\frac{1}{2}\lambda}{\frac{n}{N}\lambda + (1 - \frac{n}{N})\frac{1}{2}}.$$

When n is close to N, θ can be very close to 1. When banks conduct credit worthiness testing, lending standards (loosely defined) can affect lending in two ways. First, those applicants that were tested can be rejected if banks find them to be bad types; second, those applicants that were not tested can be rejected if the proportion of applicants that are tested is large. The second "rejected" category might contain some good type applicants. Therefore, some non-tested applicants cannot get loans if both banks test a large portion of all applicants. This is a "credit crunch" in which applicants not tested by either bank are denied loans, even if they are in fact good types.

The above discussions lead to our empirical tests in the next section: banks' relative performance is important for the credit cycles, which have a significant impact on the economy. In normal periods, banks produce information about borrowers at the optimal level, and they trigger the punishment phase by over-producing information after observing an abnormal difference in loan performance. The over-production of information leads to credit crunches. More specifically, banks will observe the relative performance differences with respect to loan portfolio size and loan defaults in the portfolio. Their beliefs about the rival banks' credit standards are based on this information. Our empirical tests are based on using measures of this information as proxies for bank beliefs.

14.3. EMPIRICAL TESTS: CREDIT CARD LOANS

In the model banks form beliefs based on public information. While we cannot measure beliefs directly, we can measure the information used to form beliefs. Our measures are proxies for bank beliefs. The empirical strategy we adopt is to focus on one robust prediction that the theory puts forward, namely, that unlike a perfectly competitive lending market, in the imperfectly competitive lending market that we have described, public histories about rival banks should affect the decisions of any given bank. We construct measures of the relative performance histories of banks, variables that are at the root of beliefs and their

formation. In particular, changes in beliefs about rival behavior should be a function of bank public performance differences.

In the U.S. the most important public information available about bank performance is the information collected by U.S. bank regulatory authorities (the Federal Reserve, Federal Deposit Insurance Corporation, and the Office of the Comptroller of Currency) in the quarterly *Call Reports of Condition and Income* (*"Call Reports"*). While publicly-traded banks also file with the Securities and Exchange Commission, the *Call Reports* provide the detail on specific loan category amounts outstanding, charge-offs, and losses. We construct Performance Difference Indices (*PDI*) based on the *Call Reports* that U.S. banks file quarterly to bank regulators. These reports are filed by banks within 30 days after the last business day of the quarter, and become public roughly 25 to 30 days later.[16] For that reason, we try to use more than one lag when we analyze the predictive power of certain variables to be constructed based on the *Call Reports*. Because the reports appear at a quarterly frequency, we analyze data at that frequency.

To parameterize the relative bank performance for our empirical studies, we use the absolute value of performance differences. Taking the absolute value is motivated by the theory. Even if a bank is doing relatively better than its rivals, it knows that if rivals believe that it has deviated then they will increase their information production, causing the better performing bank to also raise its information production. Banks, whether relatively better performing or relatively worse performing, punish simultaneously, resulting in the credit crunch. If banks' beliefs about rivals' actions change based on our parameterization of the public history, then when this measure increases, i.e., when there is a greater dispersion of relative performance, then all rival banks reduce their lending and increase its quality, resulting in fewer loans, lower loss ratios, and reduced profitability in the future. We construct indices of the absolute value of the difference in loan loss ratios and test whether the histories of such variables have predictive power for future lending decisions, loan losses, and bank stock returns.

Another challenge for testing concerns identifying rival banks. We must identify banks that are, in fact, rivals in a lending market. It is not clear whether banks compete with each other in all lending activities or only in some specialized lending areas. It is also not clear whether bank competition is a function of geography or possibly bank size. These are empirical issues.

16. Today banks submit their Call Reports electronically to Electronic Data Systems Corporation. It is then sent to the Federal Reserve Board and to the Federal Deposit Insurance Corporation, which subsequently release the data. This has of course changed over time. Nowadays, the information is available 25–30 days after it is filed on the web. Earlier private information providers would obtain computer tapes of the information from the National Technical Information Service of the Department of Commerce. The information was then provided in published formats. We thank Mary West of the Federal Reserve Board for information on the timing of the reports.

While the model suggests that there are two "regimes," normal times and punishment times, this is an artifact of simplifying the model. There could be a range of punishments, making the notion of a "regime" less discontinuous. This too is an empirical issue.

14.3.1. The Credit Card Loan Market

We first examine a specific, but important category of loans, credit card loans.[17] In the U.S. credit card lending market, potential rival banks are identifiable because credit card lending is highly concentrated and this concentration has been persistent. The Federal Reserve has collected data on credit card lending and related charge-offs since the first quarter of 1991 in the Call Reports.

The data we use is at the bank holding company level, as aggregated by the Federal Reserve Bank of Chicago. Thus, we are thinking of banks competing at the holding company level rather than at the individual bank level. For each bank holding company, we collect quarterly data from 1991.I through 2006.III for "Credit Cards and Related Plans," as well as some other variables discussed below.[18]

The high concentration is shown by the Herfindahl Index for bank holding companies as well as the market share of top bank holding companies in Figure 14.2.

We can see from Figure 14.2 that over time the credit card loan market has become increasingly concentrated; the Herfindahl Index and the market share of the top bank holding companies have become much larger.

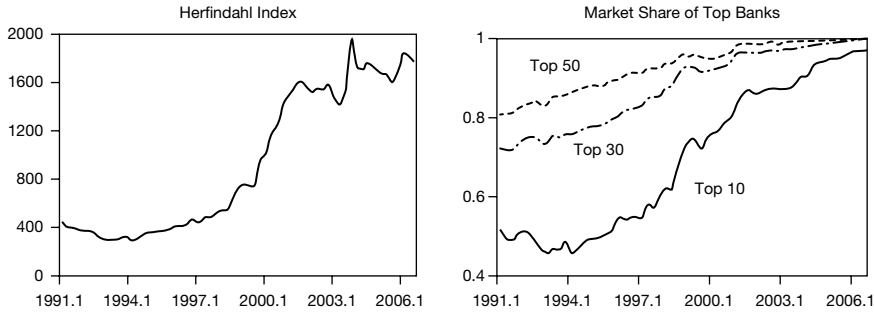


Figure 14.2 Market Concentration in Credit Card Loan Market

17. Despite the public availability of credit scores on individual consumers, banks retain important private information about credit card borrowers. Gross and Souleles (2002) show the additional explanatory power of private internal bank information in predicting consumer defaults on credit card accounts, using a sample where they were able to procure the private information.

18. The data are not reported more frequently than quarterly.

14.3.2. Data Description

The basic idea of the first set of tests is to regress an individual bank's credit card loans outstanding, normalized by total loans, or the bank's (normalized) credit card loss rate, on lagged variables that we hypothesize predict the bank's decision to make more credit card loans or to reduce losses on credit card loans (by making fewer loans or more high quality loans). Macroeconomic variables that characterize the state of the business cycle are one set of predictors. Lagged measures of the bank's own performance in the credit card market are another set of predictors. The key variables are measures of rival banks' relative histories that we hypothesize are the basis for each bank's beliefs about whether rivals have deviated. Our hypothesis is that these measures of bank histories will be significantly negative, even conditional on all the other variables.

In addition to collecting the quarterly bank holding company data from 1991.I to 2006.IV for "Credit Cards and Related Plans (LS)," we also use "Charge-offs on Loans to Individuals for Household, Family, and Other Personal Expenditure—Credit Cards and Related Plans (CO)," "Recoveries on Loans to Individuals for Household, Family, and Other Personal Expenditures—Credit Cards and Related Plans (RV)," and "Total Loans and Leases, Net (TL)." We construct the following variables for each bank holding company at quarterly level:

$$\text{Credit Card Loan Loss Ratio } (LL) = (CO - RV)/LS$$
$$\text{Ratio of Credit Card Loans to Total Loans } (LR) = LS/TL.^{19}$$

With respect to macroeconomic data we use quarterly macroeconomic data from the Federal Reserve Bank of St. Louis for the period 1991.I to 2006.III: "Civilian Unemployment Rate, Percent, Seasonally Adjusted (UMP)," "Real Disposable Personal Income, Billions of Chained 1996 Dollars, Seasonally Adjusted Annual Rate (DPI)," "Federal Funds Rate, Averages of Daily Figures, Percent (FFR)."[20]

19. Before 2001, there are two categories in Consumer Loans: Credit Card Loans & Related Plans and Other Consumer Loans. Since 2001, there are three categories in Consumer Loans: (i) Credit Card Loans, (ii) Other Revolving Credit Plans, and (iii) Other Consumer Loans. However, since 2001, the loan loss information (charge-offs and recoveries) is reported in two categories, for (i) and (ii) + (iii) respectively. Starting from 2001, we construct Loan Loss Ratio (LL) with information on Credit Card Loans only, while the Credit Card Loan Ratio (LR) is constructed using Credit Card Loans and Other Revolving Credit Plans to be consistent with before 2001.

20. We collected the monthly data for the Unemployment Rate (UMP), Disposable Income (DPI), Federal Funds Rate (FFR), and calculated the three-month averages to get the quarterly data. Also, DPI is normalized by GDP.

14.3.3. Pairwise Tests of Rival Banks

We start by looking at banks pairwise. We do this for two reasons. First, it is not known which banks are rivals, and it may be that not all banks are rivals despite the fact that they are all major credit card lenders.[21] Second, we only have less than 60 quarterly observations for each bank, so examining several banks jointly (including lags of each individual bank's performance) quickly uses up the degrees of freedom. We focus on the largest six bank holding companies, which constantly remain within the top 20 in credit card loan portfolio size during the period 1991.I to 2004.II.[22] These six banks are: JP Morgan Chase, New York, NY (CHAS); Citicorp, New York, NY (CITI); Bank One Corp., Chicago, IL (BONE); Bank of America, Charlotte, NC (BOAM); MBNA Corp., Wilmington, DE (MBNA); and Wachovia Corp., Winston-Salem, NC (WACH).

In general, we run the following regression for each bank holding company i:

$$y_{it} = \alpha_{ij} x_{it} + \beta_{ij} z_{ijt} + \varepsilon_{ijt}, \text{ for } j \neq i, \tag{14.1}$$

where

$$y_{it} = LL_{it} \text{ or } LR_{it},\ x_{it} = (Const.,\ DPI_t,\ UMP_t,\ LL_{it-1},\ LL_{it-2},\ LL_{it-3},\ LL_{it-4}),$$
$$z_{ijt} = \left(\left|\Delta LL_{ijt-1}\right|, \left|\Delta LL_{ijt-2}\right|, \left|\Delta LL_{ijt-3}\right|, \left|\Delta LL_{ijt-4}\right|\right),$$

and α_{ij} and β_{ij} are the coefficients for x and z, respectively. Adding lags of *DPI* or *UMP* do not change our major results. Since some bank holding companies might have systematically higher (or lower) loan loss rates than another bank holding companies, we first take out the mean from the loan loss ratio of each bank, and then take the difference to get ΔLL_{ij}. In this way, $|\Delta LL_{ij}|$ reflects the relative performance of the two banks.

$|\Delta LL_{ij}|$ is the key variable. It is a particular parameterization of the relevant public information: the performance difference. Conditional on the state of the economy and bank holding company i's own past performance, we ask whether bank holding company i's lending decisions depend on the observed absolute value of the differences between its own past performance and that of its rival, bank holding company j. Our theory predicts that, when $|\Delta LL_{ij}|$ and its lags are large, the bank will (implicitly) raised lending standards, resulting in fewer loans in the future and lower losses per dollar loaned. So, the coefficients are predicted to be negative. For each measure of the relative difference in loan performance,

21. For example, individual banks may dominate certain clienteles or geographical areas.

22. Data for Wachovia stops at 2001.II, as its credit card loans are managed by MBNA after that. However, the credit card loans from Wachovia do not appear in MBNA's balance sheet. After 2004.II, Bank One is acquired by JP Morgan Chase, so we do not use the data after that.

we test whether the vector of coefficients on z_{ijt} (the β's) is zero, i.e., $\beta = 0$, using a Wald test (chi-squared distribution).

An important issue with the above approach of pairwise regressions is that we do not know how many significant chi-squared statistics would be expected to be significant in a small sample. We address this issue using a bootstrap (see Horowitz (2001) for a survey). We bootstrap to test if the pairwise regression results can verify our conjecture that the measures of bank holding companies' loan performance affect each other's loan decisions. The Null hypothesis is that a bank holding company's loan decision only depends on the aggregate economic variables and its own past loan performance, i.e.:

$$H_0 : y_{it} = \alpha_i x_{it} + u_{it}.$$

The alternative hypothesis comes from the pairwise regression for each bank holding company i and bank holding company $j \neq i$:

$$H_1 : y_{it} = \alpha_{ij} x_{it} + \beta_{ij} z_{ijt} + \varepsilon_{ijt}, \text{ with } \beta_{ij} < 0.$$

In order to test the Null hypothesis, we first construct a Significance Index, *SI*, and then use the bootstrap to obtain an approximation to the distribution of the Significance Index under null hypothesis to find the p-value of the Significance Index from the pairwise regressions using the original data, *SI**. The details of the bootstrap procedure are contained in Appendix F.

The results of the pairwise regressions and the bootstraps are reported in Table 14.1. With the bootstrap we can address the question of the likelihood that adding *PDI* to the model will yield the same number of significant coefficients as with the real data. The results show that this probability is low; therefore the null hypothesis (that *PDI* is unimportant) is rejected. See the p-values for the Significance Index shown in Table 14.1.

An alternative explanation is that banks learn about the underlying economic conditions from other banks' loan performance. Perhaps this learning effect is also captured by the $|\Delta LL_{ij}|$ variable that we constructed. It would seem that learning should not be based on absolute differences in bank performance, but on the level of other banks' performances as well as the bank's own performance history. To examine this possibility we add lags of LL_j in the regression of Bank i. Therefore, in the regression equation (14.1), we replace x_{it} with x_{ijt}:

$$\begin{aligned} x_{ijt} = (&C, DPI, UMP, LL_{it-1}, LL_{it-2}, LL_{it-3}, LL_{it-4}, LL_{jt-1}, \\ &LL_{jt-2}, LL_{jt-3}, LL_{jt-4}). \end{aligned}$$

The results for learning effect are also reported in Table 14.1.

In Table 14.1, we report the average value of the coefficients on z_{ijt} as well as whether they are jointly significant. Significant negative coefficients are marked

Table 14-1. This table contains the results for pairwise regressions. In panel A and C, for each pair of banks, we run the regression: $y_{it} = \alpha_{ij} x_{it} + \beta_{ij} z_{ijt} + \varepsilon_{ijt}$, with $y_{it} = LL_{it}$ or $LR_{it}, x_{it} = (C, UMP, DPI, LL_{it-1}, LL_{it-2}, LL_{it-3}, LL_{it-4})$ and $z_{ijt} = (|\Delta LL_{it-1}|, |\Delta LL_{it-2}|, |\Delta LL_{it-3}|, |\Delta LL_{it-4}|)$. In panel B and D, for each pair of banks, we run the regression: $y_{it} = \alpha x_{ijt} + \beta_{ij} z_{ijt} + \varepsilon_{ijt}$ with $y_{it} = LL_{it}$ or $LR_{it}, x_{it} = (C, UMP, DPI, LL_{it-1}, LL_{it-2}, LL_{it-3}, LL_{it-4}, LL_{jt-1,u} LL_{jt-2}, LL_{jt-3,}, LL_{jt-4})$ and $z_{ijt} = (|\Delta LL_{it-1}|, |\Delta LL_{it-2}||\Delta LL_{it-3}|, |\Delta LL_{it-4}|)$. We report the average coefficients on z_{ijt} for each pair of banks as well as the Wald-test for the significance of these coefficients. We mark each significant average coefficient with "*" or "#" depending on the sign of the average coefficient: "*" for negative sign and "#" for positive sign. The number of "*" or "#" indicates the level of significance: Three for p-value < 0.01, two for 0.05, one for 0.10

Panel A

$y_{it} = LL_{it}$	CHAS	CITI	BONE	BOAM	MBNA	WACH
CHAS		–0.583 ***	0.064	0.044	–0.061 **	–0.446 ***
CITI	–0.175		–0.066	0.063	–0.010	–0.209 ***
BONE	–0.036	–0.246 ***		–0.228	–0.387 **	–0.302 ***
BOAM	0.307 ##	–0.127	–0.081		–0.173	0.022 ##
MBNA	0.117	–0.023	0.043	–0.054		–0.161 ***
WACH	–0.051	–0.115 ***	–0.185 *	0.096	–0.241 ***	

Significance Index: 39 Bootstrap P-Value: 0.00079

Panel B

$y_{it} = LL_{it}$	CHAS	CITI	BONE	BOAM	MBNA	WACH
CHAS		–0.641 ***	0.030	0.029	0.010	–0.231
CITI	–0.278 **		–0.122 *	0.064	0.009	–0.195 ***
BONE	–0.119 *	–0.299 ***		–0.183	–0.519 ***	–0.380 **
BOAM	0.248 #	–0.113 ***	–0.087		–0.268	0.062
MBNA	0.153 ##	–0.053 **	–0.046 ***	–0.183 **		–0.090 **
WACH	–0.061	–0.111 ***	–0.155 ***	0.029	–0.195	

Significance Index: 45 Bootstrap P–Value: 0.00001

Panel C

$y_{it} = LR_{it}$	CHAS	CITI	BONE	BOAM	MBNA	WACH
CHAS		–0.574 **	–0.077	–0.259 **	0.419	–0.010 ***
CITI	0.646		–0.590 ***	–0.572 ***	–0.224	–0.327 ***
BONE	–0.375	–0.652 ***		–1.187 ***	–0.875	–1.316 **
BOAM	–0.228	–0.497 ***	–0.184		–0.959 ***	–0.115 ***
MBNA	–0.131	0.440	0.956	0.990 ##		0.900 ###
WACH	0.475 #	–0.217	–0.439 *	0.047	–0.499 **	

Significance Index: 44 Bootstrap P-Value: 0.00011

Panel D

$y_{it} = LR_{it}$	CHAS	CITI	BONE	BOAM	MBNA	WACH
CHAS		–0.522	–0.078	–0.405 ***	0.496	–0.186
CITI	–0.074		–0.630 ***	–0.615 ***	–0.075	–0.351 *
BONE	–0.379	–0.885 ***		–1.184 ***	–1.117	–1.355
BOAM	–0.201	–0.350 ***	0.139		–0.742 ***	–0.080
MBNA	–1.515	–0.750	1.392 ###	1.324 ##		0.961 #
WACH	0.651 ##	–0.497 ***	–0.456	–0.026	–0.845 ***	

Significance Index: 38 Bootstrap P-Value: 0.00001

by '*,' and significant positive coefficients are marked by '#.' Most coefficients are negative, which matches the theoretical prediction. When the difference between the loan performance history is large, it leads to (an increase in lending standards and, consequently) a subsequent decrease in (lower quality) loans and a consequent reduction in loan losses. Many negative coefficients are significant (indicated by *** for the 1% level, by ** for the 5% level, and by * for the 10% level, and similarly for positive coefficients). Also, the Significance Indices all have very low *p*-values in our test using bootstrap.

A literal interpretation of the model would mean that there are two "regimes," rather than a possible large number of levels of intensity of information production. Perhaps there is a threshold effect, in that only if the absolute performance differences reach a certain critical level does (mutual) punishment occur. We estimated such a model using maximum likelihood and the results were not uniformly improved compared to those reported above (and so the results are omitted).

14.3.4. An Aggregate Performance Difference Index

Based on the success of the pairwise tests, we move next to analyzing the histories of all relevant rival credit card lenders jointly. We construct an aggregate Performance Difference Index (PDI):

$$PDI_t = \frac{\sum_{i>j} \left| LL_{it} - LL_{jt} \right|}{15}.$$

This Performance Difference Index measures the average difference of the competing banks' loan performances. Again, we first take out the mean from each LL_i, and then take the difference. For each bank i, we estimate the following model:

$$y_{it} = \alpha_i x_{it} + \beta_i z_t + \varepsilon_{it},\ i = 1, \ldots, 6, \qquad (\mathbf{14.2})$$

where y_{it} and x_{it} are the same as in regression (14.1), and $z_t = (PDI_{t-1}, PDI_{t-2}, PDI_{t-3}, PDI_{t-4})$. The coefficients on z_t and their t-statistics are reported in Table 14.2.

In a more restrictive environment, we estimate a pooling regression model with the restriction $\beta_i = \beta$ for $i = 1, \ldots, 6$. The results are also reported in Table 14.2.

From Panel A and C in Table 14.2, we observe that most coefficients are negative, consistent with our conjecture from the theory. When there is a large performance difference across all the rival banks, banks raise their lending standards to punish each other, and consequently future loan losses and loan ratios go down. In particular, in regressions with $y_{it} = LL_{it}$, the coefficients for JP

Table 14-2. THIS TABLE CONTAINS THE RESULTS FOR PERFORMANCE DIFFERENCE INDEX (*PDI*) REGRESSIONS. IN PANEL A AND C, FOR EACH BANK, WE RUN THE REGRESSION: $y_{it} = \alpha_i x_{it} + \beta_i z_t + \varepsilon_{it}$, WITH $y_{it} = LL_{it}$ OR LR_{it}, $x_{it} = (C, UMP, DPI, LL_{it-1}, LL_{it-2}, LL_{it-3}, LL_{it-4})$ AND $z_t = (PDI_{t-1}, PDI_{t-2}, PDI_{t-3}, PDI_{t-4})$. IN PANEL B AND D, WE POOL THE DATA OF SIX BANKS TOGETHER AND ESTIMATE THE SYSTEM WITH THE RESTRICTION THAT β_iS ARE THE SAME ACROSS BANKS: $y_{it} = a_i x_{it} + \beta z_t + \varepsilon_{it}$, WITH $y_{it} = LL_{it}$ OR LR_{it}, $x_{it} = (C, UMP, DPI, LL_{it-1} LL_{it-2}, LL_{it-3}, LL_{it-4})$ AND $z_t = (PDI_{t-1} PDI_{t-2}, PDI_{t-3}, PDI_{t-4})$ FOR $i = 1, \ldots, 6$. THE SYSTEM IS ESTIMATED USING ORDINARY LEAST SQUARES (OLS) AND SEEMINGLY UNRELATED REGRESSION (SUR) METHODS. WE REPORT THE COEFFICIENTS ON z_t AS WELL AS THEIR *t*-STATISTICS.

Panel A

	CHAS		**CITI**		**BONE**		**BOAM**		**MBNA**		**WACH**		**Panel B: Pooled**			
													OLS		**SUR**	
$y_{it} = LL_{it}$	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**
PDI_{t-1}	−0.942	−2.10	−0.279	−0.50	−1.392	−1.42	−1.380	−2.48	−0.089	−0.47	−0.679	−3.13	−0.818	−4.30	−0.563	−4.38
PDI_{t-2}	0.039	0.09	0.140	0.27	−0.786	−0.81	−0.040	−0.07	0.080	0.41	−0.393	−1.65	−0.169	−0.88	−0.202	−1.51
PDI_{t-3}	0.161	0.35	0.161	0.31	0.135	0.14	0.099	0.17	−0.005	−0.03	−0.048	−0.20	−0.028	−0.14	−0.017	−0.13
PDI_{t-4}	−0.098	−0.22	−0.117	−0.24	−1.100	−1.19	−0.453	−0.75	0.095	0.53	−0.546	−2.31	−0.341	−1.81	−0.036	−0.27
R^2	0.77		0.75		0.83		0.71		0.88		0.83					

Panel C

	CHAS		**CITI**		**BONE**		**BOAM**		**MBNA**		**WACH**		**Panel D: Pooled**			
													OLS		**SUR**	
$y_{it} = LR_{it}$	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**
PDI_{t-1}	0.144	0.30	−1.746	−2.48	−1.880	−0.78	−0.710	−2.45	0.616	0.52	0.933	0.12	−0.535	−1.32	−0.403	−2.23
PDI_{t-2}	−0.068	−0.14	−1.407	−2.16	−3.784	−1.58	−0.386	−1.33	−0.353	−0.29	−0.498	−0.58	−0.823	−2.02	−0.578	−3.19
PDI_{t-3}	−0.214	−0.44	−1.557	−2.40	−3.826	−1.61	−0.315	−1.04	−0.697	−0.57	−0.727	−0.83	−1.149	−2.79	−0.665	−3.57
PDI_{t-4}	0.187	0.39	−1.579	−2.60	−5.909	−2.61	−0.862	−2.74	1.030	0.92	−0.578	−0.67	−0.932	−2.32	−0.741	−4.02
R^2	0.74		0.89		0.75		0.92		0.88		0.83					

Morgan Chase, Bank of America, and Wachovia are statistically significant; in regressions with $y_{it} = LR_{it}$ the coefficients for Citicorp, Bank One, and Bank of America are statistically significant. In our pooling regressions, the significance of our Performance Difference Index is improved.

The coefficients are also economically significant. For example, in the regressions with Bank of America, the average coefficients on *PDI* are –0.444 and –0.568, for $y_{it} = LL_{it}$ and $y_{it} = LR_{it}$ respectively. The means of *LL* and *LR* are 0.0237 and 0.0579, respectively. Given that the standard deviation of *PDI* is 0.00454, when *PDI* changes by one standard deviation, *LL* decreases by 0.00202 (9% of the mean), and *LR* decreases by 0.00258 (5% of the mean). For Bank One, which has the largest absolute value in regression coefficients on *PDI,* the average coefficients on *PDI* for *LL* and *LR* are –0.786 and –3.850. The mean of *LL* and *LR* are 0.0316 and 0.0911. When *PDI* changes by one standard deviation, *LL* decreases by 0.00357 (11% of the mean), and *LR* decreases by 0.0275 (19% of the mean).

14.3.5. Bank Stock Returns and Performance Differences

In a credit crunch banks make fewer loans and spend more on information production, so their profitability declines. In this section, we test that implication of the model. Specifically, we ask whether the Performance Difference Index has predictive power for the stock returns of each top bank holding company in credit card loans. We collect the stock returns from CRSP from 1991.I to 2004.II. We carry out the tests for all six bank holding companies. According to our theory, after observing large performance differences between banks, banks will raise their lending standards (which is costly), and cut lending. Consequently, their profit margins will be lower. Therefore, we expect to see negative loadings on the lags of the *PDI*. Note that this is not an asset pricing model, but a test concerning bank profits, as measured by stock returns. The regression equations are:

$$r_{it} = \alpha_i + \beta_i z_t, \qquad i = 1, \ldots, 6, \qquad (\mathbf{14.3})$$

where $z_t = (PDI_{t-1}, PDI_{t-2}, PDI_{t-3}, PDI_{t-4})$.

Since the dividend yield is well known to be a predictor of future stock returns (see, for example, Cochrane (1999)), we also estimate the model with the lag of dividend yield as a predicting variable. Again, robustness is checked by imposing the restriction $\beta_i = \beta$ for $i = 1, \ldots, 6$. All the results are reported in Table 14.3.

From Table 14.3, we see that the *PDI* from the previous four quarters significantly predicts the stock return for the current quarter, and the results are robust if we include a lag of the dividend yield in the regressions. The average

Table 14-3. THIS TABLE CONTAINS THE RESULTS FOR THE PREDICTIVE POWER OF PERFORMANCE DIFFERENCE INDEX (*PDI*) FOR STOCK RETURNS. IN PANEL A AND C, FOR EACH BANK, WE RUN THE REGRESSION: $r_{it} = a_i x_{it} + \beta_i z_t + \varepsilon_{it}$, WITH $x_{it} = C$ OR $(C, Dividend\ Yield_{it-1})$ AND $z_{it} = (PDI_{t-1}, PDI_{t-2}, PDI_{t-3}, PDI_{t-4})$. IN PANEL B AND D, WE POOL THE DATA OF SIX BANKS TOGETHER AND ESTIMATE THE SYSTEM WITH THE RESTRICTION THAT β_iS ARE THE SAME ACROSS BANKS: $r_{it} = a_i x_{it} + \beta z_t + \varepsilon_{it}$,FOR $i = 1, \ldots, 6$. THE SYSTEM IS ESTIMATED USING ORDINARY LEAST SQUARES (OLS) AND SEEMINGLY UNRELATED REGRESSION (SUR) METHODS. WE REPORT THE COEFFICIENTS ON z_t AS WELL AS THEIR *t*-STATISTICS

Without																
Dividend	**Panel A**												**Panel B:Pooling**			
	CHAS		**CITI**		**BONE**		**BOAM**		**MBNA**		**WACH**		**OLS**		**SUR**	
Yield	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**
PDI_{t-1}	−3.66	−0.75	3.53	0.66	1.46	0.39	1.35	0.36	−8.28	−1.60	−1.66	−0.48	−1.21	−0.68	−0.82	−0.32
PDI_{t-2}	−2.56	−0.53	−1.73	−0.32	−2.53	−0.67	−3.09	−0.82	1.56	0.30	−6.76	−1.95	−2.50	−1.40	−3.77	−1.48
PDI_{t-3}	−9.78	−2.02	−4.80	−0.90	−8.91	−2.37	−9.97	−2.63	−6.87	−1.33	2.00	0.57	−6.45	−3.59	−4.66	−1.83
PDI_{t-4}	−1.60	−0.32	−4.97	−0.91	−7.13	−1.86	−5.91	−1.53	−4.71	−0.90	−5.80	−1.65	−5.04	−2.77	−6.24	−2.41
R^2	0.13		0.07		0.14		0.25		0.13		0.16					
With																
Dividend	**Panel C**												**Panel D: Pooling**			
	CHAS		**CITI**		**BONE**		**BOAM**		**MBNA**		**WACH**		**OLS**		**SUR**	
Yield	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**	**Coeff.**	***t*-stat**
PDI_{t-1}	−2.72	−0.54	4.62	0.80	4.38	1.28	1.13	0.29	−8.20	−1.56	−2.71	−0.74	−0.83	−0.45	−0.69	−0.28
PDI_{t-2}	−1.65	−0.33	−1.97	−0.36	−0.66	−0.20	−3.20	−0.84	1.65	0.31	−6.96	−2.00	−2.15	−1.20	−3.48	−1.41
PDI_{t-3}	−8.72	−1.72	−4.33	−0.79	−7.47	−2.24	−9.72	−2.53	−6.81	−1.30	0.59	0.15	−6.28	−3.43	−5.33	−2.13
PDI_{t-4}	−0.41	−0.08	−4.55	−0.82	−3.38	−0.95	−5.63	−1.44	−4.64	−0.87	−5.54	−1.56	−4.13	−2.23	−5.35	−2.12
R^2	0.14		0.08		0.43		0.26		0.13		0.18					

coefficient on the lags of *PDI* from OLS estimates is about –3.5. One standard deviation change in *PDI* (0.00454) leads to an average change of 0.0159 in stock returns, or 159 basis points!

14.3.6. Rajan's Reputation Hypothesis

Rajan (1994) argues that reputation considerations of bank managers cause banks to simultaneously raise their lending standards when there is an aggregate shock to the economy causing the loan performance of all banks to deteriorate. Banks tend to neglect their own loan performance history in order to herd or pool with other banks. Rajan's empirical work focuses on seven New England banks over the period 1986–1991. His main finding is that a bank's loan charge-offs-to-assets ratio is significantly related not only to its own loan loss provisions-to-total assets ratio, but also to the average charge-offs-to-assets ratio for other banks (instrumented for by the previous quarter's charge-offs-to-assets ratio).[23] In the context here the question is whether our measure of banks' beliefs about rivals' credit standards, the Performance Difference Index, remains significant in the presence of an average or aggregate credit card loss measure. We construct:

$$\text{Aggregate Credit Card Loan Loss } (AGLL_t) = \frac{\sum_i (CO_{it} - RV_{it})}{\sum_i LS_{it}}$$

and then examine the coefficients on the lags of *AGLL* and *PDI,* separately and jointly, in our regression equation (14.2) with $z_t = (AGLL_{t-1}, AGLL_{t-2}, AGLL_{t-3}, AGLL_{t-4})$ or $z_t = (AGLL_{t-1}, AGLL_{t-2}, AGLL_{t-3}, AGLL_{t-4}; PDI_{t-1}, PDI_{t-2}, PDI_{t-3}, PDI_{t-4})$.

The coefficients on z_t and their *t*-statistics are reported in Table 14.4, which also contains the results with the restriction that the coefficients on z_t are the same across bank holding companies.

Rajan's (1994) hypothesis is that an aggregate bad shock leads banks to raise their standards, so we would expect the coefficients on lags of *AGLL* to be significantly negative. However, as Table 14.4 shows, with or without *PDI* in the regressions, the coefficients on *AGLL* are mostly positive and significant, with a few exceptions. At the same time, the coefficients on lags of *PDI* remain negatively significant, even after we include lags of *AGLL* in our regression.

23. There are several interpretations of Rajan's result. For example, the charge-offs of other banks may be informative about the state of the economy, so their significance in the regression is not necessarily evidence in favor of Rajan's theory.

Table 14-4. THIS TABLE CONTAINS THE RESULTS OF TESTING RAJAN'S (1994) REPUTATION HYPOTHESIS. IN PANEL A AND C, WE POOL THE DATA OF SIX BANKS TOGETHER AND ESTIMATE THE SYSTEM: $y_{it} = \alpha_i x_{it} + \beta z_t + \varepsilon_{it}$, WITH $y_{it} = LL_{it}$ OR $LR_{it}, x_{it} = (C, UMP, DPI, LL_{it-1}, LL_{it-2}, LL_{it-3}, LL_{it-4})$ AND $z_t = (AGLL_{t-1}, AGLL_{t-2}, AGLL_{t-3}, AGLL_{t-4})$ for $i = 1,\ldots,6$. IN PANEL B AND D, WE POOL THE DATA OF SIX BANKS TOGETHER AND ESTIMATE THE SYSTEM: $y_{it} = \alpha_i x_{it} + \beta z_t + \varepsilon_{it}$, WITH $y_{it} = LL_{it}$ OR $LR_{it}, x_{it} = (C, UMP, DPI, LL_{it-1}, LL_{it-2}, L_{it-3}, LL_{it-4})$ AND $z_t, = (AGLL_{t-1}, AGLL_{t-2}, AGLL_{t-3}, AGLL_{t-4}, PDI_{t-1}, PDI_{t-2}, PDI_{t-3}, PDI_{t-4})$ for $i = 1,\ldots,6$. THE SYSTEM IS ESTIMATED USING ORDINARY LEAST SQUARES (OLS) AND SEEMINGLY UNRELATED REGRESSION (SUR) METHODS. WE REPORT THE COEFFICIENTS ON z_t AS WELL AS THEIR t-STATISTICS

$y_{it} = LL_{it}$	Panel A				Panel B			
	OLS		SUR		OLS		SUR	
	Coeff.	*t*-stat	Coeff.	*t*-stat	Coeff.	*t*-stat	Coeff.	*t*-stat
$AGLL_{t-1}$	−0.023	−0.40	−0.103	−1.82	0.085	1.46	0.036	0.66
$AGLL_{t-2}$	−0.038	−0.64	−0.059	−1.03	0.110	1.84	0.088	1.56
$AGLL_{t-3}$	0.097	1.65	0.028	0.49	0.212	3.47	0.145	2.48
$AGLL_{t-4}$	0.323	5.66	0.265	4.77	0.316	5.61	0.263	4.86
PDI_{t-1}					−0.892	−5.26	−0.895	−5.70
PDI_{t-2}					−0.433	−2.40	−0.334	−2.01
PDI_{t-3}					−0.312	−1.72	−0.296	−1.77
PDI_{t-4}					−0.391	−2.25	−0.100	−0.63
$y_{it} = LR_{it}$	Panel C				Panel D			
	OLS		SUR		OLS		SUR	
	Coeff.	*t*-stat	Coeff.	*t*-stat	Coeff.	*t*-stat	Coeff.	*t*-stat
$AGLL_{t-1}$	0.102	0.51	0.043	0.38	0.094	0.49	0.080	1.01
$AGLL_{t-2}$	0.196	0.98	0.166	1.47	0.212	1.09	0.187	2.33
$AGLL_{t-3}$	0.228	1.14	0.131	1.15	0.247	1.23	0.172	2.09
$AGLL_{t-4}$	0.340	1.75	0.267	2.44	0.313	1.69	0.211	2.78
PDI_{t-1}					−0.192	−0.35	−0.222	−0.91
PDI_{t-2}					−0.797	−1.35	−0.674	−2.57
PDI_{t-3}					−1.165	−1.96	−0.835	−3.19
PDI_{t-4}					−0.817	−1.44	−0.549	−2.15

14.4. AN AGGREGATE PERFORMANCE DIFFERENCE INDEX FOR COMMERCIAL AND INDUSTRIAL LOANS

In this section we extend the empirical analysis beyond credit card lending at six banks to examine commercial and industrial loan market at an aggregate level, and we probe the implications of the theory for macroeconomic dynamics. Commercial and industrial loans is the category of loans that covers lending to firms of all sizes and corresponds to the loans at issue when there is a credit

crunch. If banks increase their information production, that is, raise their lending standards, then some borrowers are cut off from credit—a credit crunch that should have macroeconomic implications. We examine this with a vector autoregression in the first subsection. In the second subsection, we examine the Performance Difference Index less formally to get a feel for what it measures.

14.4.1. VAR Analysis of the Fed's Lending Standards Index

In this subsection, we use Vector Autoregressions (VARs) to analyze the aggregate implications of banks' loan performance differences. In contrast to the single equations estimated above, a VAR system of equations lets us control for the feedback between current and past levels of performance differences, the lending standard survey results, and macroeconomic variables. Given estimates of these interactions, we can identify the impact that unpredictable shocks in performance difference public histories have on other variables in the system. We first ask whether the performance difference histories predict, in the sense of Granger causality, the Index of Lending Standards based on the Federal Reserve System's Senior Loan Officer Opinion Survey on Bank Lending Practices. The Federal Reserve System's Senior Loan Officer Opinion Survey started in 1967.I, but was discontinued during the period 1984.I to 1990.I.

We follow Lown and Morgan (2005, 2002) in analyzing the time series of lending survey responses, the net percentage of banks reporting tightening in the survey.[24] As above, we use quarterly commercial and industrial loan data from the Chicago Federal Reserve Bank's Commercial Bank Database, which is from the *Call Reports*. For the period from 1984.I to 2006.III, we collected "Commercial and Industrial Loans to U.S. Addressees" (LS), "Charge-Offs on Commercial and Industrial Loans to U.S. Addressees" (CO), and "Recoveries on Commercial and Industrial Loans to U.S. Addressees" (RV). For each commercial bank we constructed the

$$\text{Loan Loss Ratio}\,(LL): LL = \frac{(CO - RV)}{LS}.$$

We construct the Performance Difference Index to measure the dispersion of performance across the U.S. banking industry as a whole. To do this, we use

24. Following Lown and Morgan (2005, 2001) we use the standards for large and middle-market firms. As mentioned, the Lending Standard Index is calculated as the net percentage of banks (all respondents) that report tightening.

the top 100 commercial banks[25] ranked by commercial and industrial (C&I) loans, and for each quarter, we construct the Performance Difference Index as: $PDI_t = \frac{\sum_{i>j} |LL_{it} - LL_{jt}|}{100 \times 99/2}$. Besides the data on the Lending Standards and the Performance Difference Index, we also collected data on Commercial and Industrial Loans at "All Commercial Banks and Federal Funds Rate" from the FRED II database of the St. Louis Fed.[26] As before, we conjecture that this *PDI* captures the relevant history that is at the basis of banks' beliefs about whether other banks are deviating to using the credit worthiness tests.

The VAR includes four lags of the four endogenous variables: Bank Lending Standards (*STAND*) (i.e., the net percentage of survey respondents reporting tightening), the Performance Difference Index *(PDI),* the Federal Funds Rate (*FFR*), and the log of Commercial Bank C&I Loans (*LOGLOAN*). The bank Lending Standard variable is a loan supply side factor and the Federal Funds Rate affects loan demand; Commercial Bank C&I Loan is the equilibrium outcome. The *PDI* is hypothesized to capture banks' beliefs, which affects all the other variables. The exogenous variables include a constant and a time trend. We run the VAR for the period of 1990.II–2006.III, which is the longest continuous of period where we have both *STAND* and *PDI* data. During this period of time, the means and standard deviations of these four variables are:

	STAND	***PDI***	***FFR***	***LOGLOAN***
Mean	5.572	0.00411	5.065	7.901
STD	21.311	0.00319	1.349	0.228

We report the VAR results in Table 14.5.

Table 14.5 shows that the *PDI* Granger-causes the other three endogenous variables, and only *STAND* Granger-causes *PDI* (actually none of the individual coefficients on *STAND* are significant, but they are jointly significant). For each of the other three endogenous variables, using the average coefficients on the lags of *PDI*, a one standard deviation increase in *PDI* leads to a 2.6% increase in net percentage of loan officers who claim to be raising the lending standards, an 80 basis point decrease in the federal funds rate, and a 0.44% decrease in C&I loans.

At the same time, the lending standards are significantly affected by *PDI* and *LOGLOAN*. A high level of performance differences causes a rise in lending standards, consistent with our theory of information production competition. Besides *PDI*, both *STAND* and *FFR* Granger-cause *LOGLOAN*. To further

25. We also construct the performance difference indices using top 50 or top 200 commercial banks ranked by their C&I loan size; the results are similar.

26. We first collected monthly data and then took the three-month average to obtain quarterly data.

Table 14-5. THIS TABLE PRESENTS THE AVERAGE VALUE OF THE COEFFICIENTS AND *p*-VALUES (IN PARENTHESIS) OF THE WALD TEST ($\chi^2(4)$) OF THE VAR WITH FOUR LAGS OF THE LENDING STANDARD (*STAND*), THE PERFORMANCE DIFFERENCE INDEX (*PDI*), THE FEDERAL FUNDS RATE (*FFR*), AND THE LOG OF COMMERCIAL BANK C&I LOAN (*LOGLOAN*). THE EXOGENOUS VARIABLES INCLUDE A CONSTANT AND A TIME TREND

	STAND	**PDI**	**FFR**	**LOGLOAN**
STAND	1.15E–01 (0.002)	2.19E–05 (0.004)	4.59E–04 (0.878)	–6.51E–05 (0.118)
PDI	8.10E + 02 (0.037)	2.41E–01 (0.000)	–2.51E + 01 (0.064)	–1.37E + 00 (0.000)
FFR	1.70E–01 (0.315)	6.70E–05 (0.417)	2.01E–01 (0.000)	1.83E–03 (0.000)
LOGLOAN	2.52E + 01 (0.044)	–6.27E–04 (0.416)	7.31 E–02 (0.545)	2.39E–01 (0.000)

explore the impact of *PDI* on other endogenous variables, we also report the forecasting error variance decomposition of our VAR in Table 14.6.

As we can see from Table 14.6, at a five-quarter horizon, innovations in *STAND* account for 13.9% of the error variance in the federal funds rate and 14.1% of the *LOGLOAN* error variance, while those numbers for *PDI* are 21.3% and 34.6%, respectively. At longer horizons, ten quarters and fifteen quarters, *PDI* continues to dominate *STAND* as a major variance contributor for *FFR* and *LOGLOAN*. Therefore, the Performance Difference Index has a bigger impact than Lending Standards despite the fact that in our VAR the Lending Standards variable is ranked before the Performance Difference Index variable. This confirms our view that *PDI* is a major economic indicator for bank competition, consistent with our information-based theory.

14.4.2. Understanding the Performance Difference Index

We can understand the Performance Difference Index more intuitively by noting that a higher *PDI* is bad news for consumers, since credit lending standards will become more stringent and credit card loans will go down. This would apply also to other types of consumer loans, such as home equity loans, home improvement loans, automobile and boat loans, and so on. And it is bad news for firms, especially small firms, because lending standards will be raised, making commercial and industrial loans harder to obtain.

These broad implications are confirmed in Figure 14.3 below. The figure shows plots of the year-on-year change in U.S. GDP, the Michigan Consumer Confidence Index, and the four quarter moving average of *PDI* (based on C&I loans). At business cycle peaks, Consumer Confidence declines,

Table 14-6. This table reports the results of forecasting errors and their variance decomposition among four endogenous variables. For each panel, the first column lists the number of quarters for forecasting, the second column contains the standard errors of forecasting errors for certain forecasting horizon, and the next four columns are the weight (in percentage) of each endogenous variable in contributing to the forecasting errors

Variance Decomposition of *STAND*

Period	St. Error	*STAND*	*PDI*	*FFR*	*LOGLOAN*
1	6.64	100.0	0.0	0.0	0.0
3	8.03	89.5	2.9	6.9	0.8
5	9.53	65.6	2.7	27.7	4.0
10	11.13	50.0	13.6	32.6	3.8
15	12.49	45.4	18.0	33.3	3.3

Variance Decomposition of *PDI*

Period	St. Error	*STAND*	*PDI*	*FFR*	*LOGLOAN*
1	0.00094	0.2	99.8	0.0	0.0
3	0.00101	6.0	86.2	4.8	3.0
5	0.00134	14.5	76.6	6.5	2.4
10	0.00170	14.7	57.5	24.2	3.6
15	0.00188	14.1	58.6	23.2	4.1

Variance Decomposition of *FFR*

Period	St. Error	*STAND*	*PDI*	*FFR*	*LOGLOAN*
1	0.231	10.7	4.7	84.6	0.0
3	0.527	10.6	12.5	76.1	0.8
5	0.692	13.9	21.3	64.2	0.5
10	0.869	12.9	27.2	57.6	2.3
15	1.017	11.6	20.7	65.2	2.6

Variance Decomposition of *LOGLOAN*

Period	St. Error	*STAND*	*PDI*	*FFR*	*LOGLOAN*
1	0.0050	23.8	0.5	8.1	67.7
3	0.0131	6.6	17.9	51.8	23.7
5	0.0212	14.1	34.6	40.8	10.5
10	0.0363	23.5	50.7	21.7	4.1
15	0.0519	25.1	32.4	39.0	3.5

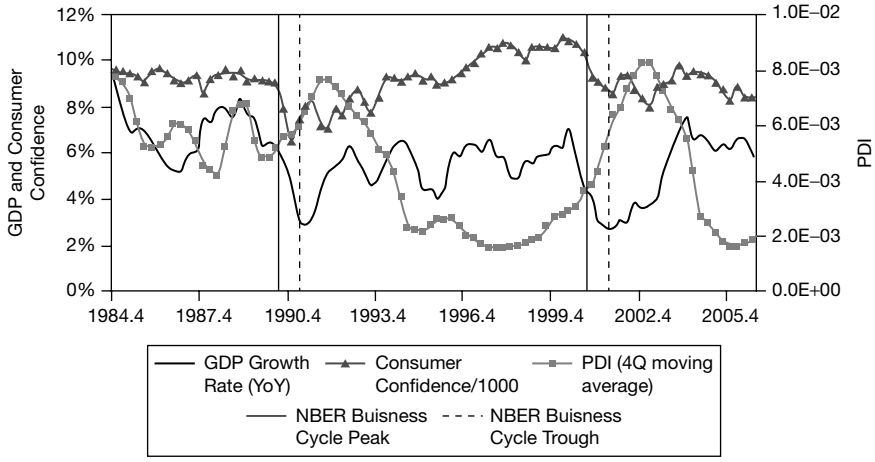


Figure 14.3 PDI, Consumer Confidence, and GDP Growth

and the year-over-year growth rate of GDP is going down. Notably, *PDI* is rising.

These observations mean the *PDI* should be negatively correlated with Consumer Confidence (as measured by the University of Michigan Survey Research Center[27]) and *PDI* should be negatively correlated with aggregate economic activity. The table below shows the relevant correlations. ("YoY" means year-over-year.)

Correlation Matrix (1984.I–2006.III)	**ΔGDP**	**CC**	**LS**	**PDI**
GDP YoY Growth Rate (AGDP)	1.00	0.33	–0.57	–0.37
Consumer Confidence (CC)	0.33	1.00	–0.09	–0.47
Lending Standards starting from 1990.II (LS)	–0.57	–0.09	1.00	0.46
Performance Difference Index (PDI) (Deseasoned)	–0.37	–0.47	0.46	1.00

As expected, *PDI* is negatively correlated with Consumer Confidence and with the year-on-year GDP growth rate. As noted in the VAR analysis, *PDI* and Lending Standards are positively correlated.[28]

27. See http://www.isr.umich.edu/src/.

28. The credit card market and the commercial and industrial loan market need not display credit crunches at the same time, as banks may behave as if they are separate markets. The two PDI indices for these markets have a correlation of 0.18 after being deseasoned, and 0.46 before being deseasoned.

14.5. ASSET PRICING AND CREDIT CRUNCHES

Strategic competition between banks results in periodic credit crunches, which are a systematic risk even though endogenous. Consequently, if the stock market is efficient, then the stock returns of both banks and non-financial firms, which, at least partially, rely on banks for external financing, should reflect the competition between banks. In this section we turn to a different empirical approach, namely, we look for the hypothesized systematic effects in an asset pricing context.

If strategic behavior between banks causes credit cycles, then it causes variation in the profitability of non-financial firms. Credit crunches are also not profitable for banks. The credit cycle is a systematic risk (even if it is endogenous, emanating from bank competition) and therefore should be a priced factor in stock returns, to the extent that this factor is not already spanned by other factors. We conjecture that the constructed *PDI* should be a priced risk factor for both banks and non-financial firms. That is, in the context of an asset pricing model of stock returns, there should be an additional factor, namely, the Performance Difference Index. Moreover, since relatively smaller firms are more dependent on bank loans (see, e.g., Hancock and Wilcox (1998)), we expect that the coefficients on *PDI* (below, we construct the mimicking portfolio for this factor) are larger for smaller firms.

We adopt the classic Capital Asset Pricing Model as the benchmark for examining whether *PDI* is a priced factor. Later, we will also examine the Fama-French three factor empirical asset pricing model.[29] The model is estimated using quarterly data, as *PDI* can only be calculated quarterly.

We hypothesize that bank stock returns will be sensitive to *PDI* and that *PDI* is not spanned by the market factor. Further, non-financial firms' stock returns will also be sensitive, increasingly so for smaller firms, to *PDI.* The monthly firm returns are collected from CRSP (then transformed into quarterly data). We separate out commercial banks and non-financial firms based on their *SIC* codes, and then divide the non-financial firms into ten deciles based on the capitalizations. Banks are divided into small, medium, and large. The data used are from 1984.I to 2006.III, during which the performance difference index is available.

As is standard in the asset pricing literature, we proceed by first constructing the mimicking portfolio for our macro factor, *PDI.* Mimicking portfolios are needed to identify the factor risk premiums when the factors are not traded

29. See Fama and French (1993, 1996). Carhart (1997) introduced an additional factor, the momentum factor. The results with the additional momentum factor are basically the same, and are thus omitted. We collected the quarterly Fama-French three factors from French website (the construction method can also be found there). The risk free rates are three-month T-Bill rates (secondary market rates) from FRED II (we use the rate of the first month in each quarter) at Federal Reserve Bank at St. Louis.

assets. The risk premium is constructed as a "mimicking portfolio" return whose conditional expectation is an estimate of the risk premium or price of risk for that factor. We then use a time series regression approach, as in, for example, Breeden, Gibbons, and Litzenberger (1989), with the book-to-market sorted portfolios as the base assets. A recent study by Asgharian (2006) argues that this approach is the best for constructing mimicking portfolios for factors for which a time-series factor realization is available.

We first regress the *PDI* factor on the excess returns of the ten book-to-market sorted portfolios (either equal-weighted or value-weighted), and then construct the mimicking portfolio with the weight of each portfolio proportional to the regression coefficient on the excess return of this portfolio. Specifically, we first run the following regression:

$$PDI_t = \lambda_0 + \sum_{i=1}^{10} \lambda_i R_{it} + \varepsilon_t,$$

where R_{it} is the excess return on the base asset i at time t. The weights are constructed as follows:

$$w_i = \frac{\lambda_i}{\sum_{i=1}^{10} \lambda_i},$$

and the excess return on the mimicking portfolio is given by:

$$R_{PDI,t} = \sum_{i=1}^{10} w_i R_{it}.$$

According to Breeden et al. (1989), the asset betas measured relative to the maximum correlation portfolio are proportional to the betas measured using the true factor.

After we form the mimicking portfolio, we add it to the standard Capital Asset Pricing Model (CAPM). The results are reported in Table 14.7. The results in Table 14.7 show that the *PDI* mimicking portfolio is a significant risk factor for small non-financial firms and for all bank sizes. Note that the coefficients on R_{PDI} for smaller firms are larger, thus confirming our conjectures. This is also confirmed by the monotonicity of the improvement in R^2 with the new *PDI* factor.

In terms of the economic significance of the new *PDI* factor, the standard deviation of R_{PDI} (constructed with value-weighted book-to-market portfolios) is 15 percent (this is quite large because the mimicking portfolio involves short positions). Therefore, when R_{PDI} changes by one standard deviation, the excess return for the smallest non-financial firms changes by about 2.6 percent. As a comparison, from 1984.I to 2006.III, for Table 14.7 (CAPM), a one standard

Table 14-7. THIS TABLE REPORTS THE RESULTS FROM ESTIMATING THE AUGMENTED CAPM MODEL: $r_i - r_f = \alpha + \beta_1 (r_m - r_f) + \beta_2 R_{PDI} + \varepsilon$, WHERE R_{PDI} IS CONSTRUCTED FROM TEN BOOK-TO-MARKET PORTFOLIOS, EITHER EQUAL WEIGHTED OR VALUE WEIGHTED. WE REPORT THE COEFFICIENTS AND THEIR *t*-STATISTICS (IN PARENTHESES), R^2 OF EACH REGRESSION, AND R^2 OF THE REGRESSION WITHOUT R_{PDI} (IN PARENTHESES)

Coefficient (*t*-stat)	α	$r_m - r_t$	R_{PDI}	R^2 (R^2 *w/o* R_{PDI})	α	$r_m - r_t$	R_{PDI}	R^2 (R^2 *w/o* R_{PDI})
	Commercial Banks (using equal weighted R_{PDI})				**Commercial Banks (using value weighted R_{PDI})**			
Small	1.988	0.456	0.119	0.33	2.227	0.447	0.109	0.25
	(2.34)	(4.45)	(4.06)	(0.20)	(2.47)	(4.00)	(2.26)	(0.20)
Medium	2.131	0.589	0.118	0.40	2.388	0.584	0.102	0.32
	(2.50)	(5.75)	(4.02)	(0.28)	(2.64)	(5.21)	(2.11)	(0.28)
Large	1.970	1.003	0.041	0.65	2.045	0.999	0.040	0.64
	(2.89)	(12.22)	(1.75)	(0.64)	(2.97)	(11.67)	(1.09)	(0.64)
	Non-Financial Firms (using equal weighted R_{PDI})				**Non-Financial Firms (using value weighted R_{PDI})**			
Decile 1	4.669	1.096	0.256	0.47	5.374	1.119	0.174	0.34
(Small)	(3.22)	(6.29)	(5.14)	(0.31)	(3.33)	(5.58)	(2.00)	(0.31)
Decile 2	−0.523	1.143	0.193	0.58	0.058	1.171	0.115	0.47
	(−0.48)	(8.73)	(5.17)	(0.45)	(0.05)	(7.72)	(1.76)	(0.45)

Table 14-7. (CONTINUED)

Coefficient (t-stat)	α	$r_m - r_t$	R_{PDI}	R^2 ($R^2 w/o\ R_{PDI}$)	α	$r_m - r_t$	R_{PDI}	R^2 ($R^2 w/o\ R_{PDI}$)
	Non-Financial Firms (using equal weighted R_{PDI})				**Non-Financial Firms (using value weighted R_{PDI})**			
Decile 3	−0.662	1.230	0.141	0.64	−0.238	1.250	0.085	0.57
	(−0.71)	(10.90)	(4.39)	(0.56)	(−0.23)	(9.86)	(1.55)	(0.56)
Decile 4	−0.762	1.265	0.129	0.68	−0.340	1.291	0.066	0.62
	(−0.87)	(12.04)	(4.31)	(0.61)	(−0.36)	(10.93)	(1.30)	(0.61)
Decile 5	−0.196	1.321	0.108	0.73	0.223	1.358	0.033	0.68
	(−0.25)	(13.99)	(3.99)	(0.68)	(0.26)	(12.87)	(0.71)	(0.68)
Decile 6	0.145	1.348	0.067	0.75	0.454	1.381	0.005	0.72
	(0.20)	(15.22)	(2.67)	(0.72)	(0.59)	(14.54)	(0.13)	(0.72)
Decile 7	0.481	1.360	0.041	0.82	0.724	1.392	−0.015	0.81
	(0.82)	(19.39)	(2.04)	(0.81)	(1.22)	(18.83)	(−0.48)	(0.81)
Decile 8	0.721	1.293	0.033	0.87	0.929	1.321	−0.015	0.86
	(1.55)	(23.12)	(2.09)	(0.86)	(1.95)	(22.41)	(−0.60)	(0.86)
Decile 9	0.939	1.179	0.017	0.91	1.077	1.201	−0.018	0.91
	(2.81)	(29.29)	(1.48)	(0.91)	(3.20)	(28.73)	(−1.02)	(0.91)
Decile 10	1.113	0.982	−0.013	0.97	1.067	0.979	−0.005	0.97
(Large)	(7.09)	(51.95)	(−2.40)	(0.97)	(6.60)	(48.74)	(−0.62)	(0.97)

deviation change of market excess return, 7.9 percent, results in the excess return for the smallest non-financial firms changing by about 8.8 percent. If we use equal-weighted book-to-market portfolios to construct our mimicking portfolio, one standard deviation change of R_{PDI} results in a 7.4 percent change of the excess return of the smallest non-financial firms, which is close to the impact of market excess return.

We conclude that the competition and collusion among banks is an important risk factor for stock returns, for banks and especially for small non-financial firms. The size effect further demonstrates that the Performance Difference Index we constructed is not capturing some sort of learning effect about macroeconomic condition, which would be spanned by the other risk factors.

As a robustness check, we will also investigate the Fama-French three factor empirical asset pricing model. According to Fama and French, the sensitivity of a firm's expected stock return depends on three factors: the excess return on a broad based market portfolio, $r_m - r_f$, the difference between the return on a portfolio of small stocks and the return on a portfolio of large stocks (small minus large), *SMB*; the difference between the return on a portfolio of high book-to-market stocks and the return on a portfolio of low book-to-market stocks (high minus low), *HML*.

One concern regarding *PDI* as a macro factor is that it might have been priced into the three factors. To address that concern, we first regress the three Fama-French factors on the Performance Difference Index to see whether there is a significant correlation between them. The results are as follows:

	Coefficient on *PDI*	***t*-statistics**
$r_m - r_f$	−238.90	−0.90
SMB	−49.79	−0.29
HML	−6.98	−0.03

We can see that none of the coefficients are significant. Therefore, *PDI* is not spanned by the other factors.

After we form the mimicking portfolio, we add it to the Fama-French three-factor model. The results are reported in Table 14.8. The results in Table 14.8 also show that the *PDI* mimicking portfolio is a significant risk factor for small non-financial firms and for small banks, but not for large banks or large non-financial firms. Again, the coefficients on R_{PDI} for smaller firms are larger, as well as the improvement in R^2 with the new *PDI* factor. Also, comparing Table 14.7 and Table 14.8, we can compare the improvement of R^2 by adding our *PDI* factor with that by adding *HML* & *SMB*. For small non-financial firms, R^2 improves from 0.31 to 0.47 by adding our *PDI* factor, and it improves from 0.31 to 0.57 by adding both *HML* & *SMB,* and further to 0.65 by adding our *PDI* factor.

Table 14-8. This table reports the results from estimating the augmented Fama-French three factor model:

$$r_i - r_f = \alpha + \beta_1 \left(r_m - r_f\right) + \beta_2 SMB + \beta_3 HML + \beta_4 R_{PDI} + \varepsilon,$$

where R_{PDI} is constructed from ten book-to-market portfolios, either equal weighted or value weighted. We report the coefficients and their t-statistics (in parentheses), R^2 of each regression, and R^2 of the regression without R_{PDI} (in parentheses)

Coefficient (t-stat)	α	$r_m - r_f$	SMB	HML	R_{PDI}	$R^2 (R^2 w/o\ R_{PDI})$	α	$r_m - r_f$	SMB	HML	R_{PDI}	$R^2 (R^2 w/o\ R_{PDI})$
	Commercial Banks (using equal weighted R_{PDI})						**Commercial Banks (using value weighted R_{PDI})**					
Small	1.578	0.489	0.587	0.613	0.060	0.55	1.631	0.449	0.603	0.718	0.071	0.54
	(2.22)	(4.80)	(5.08)	(4.11)	(2.35)	(0.52)	(2.27)	(4.09)	(5.14)	(4.87)	(1.78)	(0.52)
Medium	1.689	0.639	0.622	0.592	0.058	0.61	1.760	0.609	0.644	0.690	0.059	0.59
	(2.40)	(6.35)	(5.44)	(4.02)	(2.29)	(0.58)	(2.46)	(5.60)	(5.54)	(4.72)	(1.51)	(0.58)
Large	1.564	1.188	0.463	−0.098	0.019	0.72	1.634	1.200	0.487	−0.073	−0.003	0.72
	(2.51)	(13.28)	(4.57)	(−0.75)	(0.82)	(0.72)	(2.61)	(12.57)	(4.77)	(−0.57)	(−0.08)	(0.72)
	Non-Financial Firms (using equal weighted R_{PDI})						**Non-Financial Firms (using value weighted R_{PDI})**					
Decile 1	4.940	0.629	1.657	−0.042	0.192	0.65	5.146	0.518	1.983	0.020	0.208	0.61
(Small)	(4.08)	(3.63)	(6.53)	(−0.21)	(4.37)	(0.57)	(4.04)	(2.67)	(7.60)	(0.10)	(2.96)	(0.57)
Decile 2	−0.398	0.766	1.517	0.108	0.126	0.79	−0.256	0.696	1.731	0.151	0.134	0.76
	(−0.50)	(6.76)	(9.15)	(0.84)	(4.41)	(0.74)	(−0.31)	(5.47)	(10.13)	(1.11)	(2.90)	(0.74)
Decile 3	−0.534	0.867	1.445	0.093	0.078	0.86	−0.495	0.802	1.585	0.103	0.105	0.85
	(−0.89)	(10.04)	(11.44)	(0.95)	(3.57)	(0.84)	(−0.81)	(8.59)	(12.64)	(1.03)	(3.11)	(0.84)

Coefficient (t-stat)	α	$r_m - r_f$	SMB	HML	R_{PDI}	$R^2(R^2 w/o\ R_{PDI})$	α	$r_m - r_f$	SMB	HML	R_{PDI}	$R^2(R^2 w/o\ R_{PDI})$
	Non-Financial Firms (using equal weighted R_{PDI})						**Non-Financial Firms (using value weighted R_{PDI})**					
Decile 4	−0.702	0.946	1.370	0.160	0.065	0.88	−0.649	0.901	1.483	0.175	0.078	0.88
	(−1.31)	(12.28)	(12.15)	(1.83)	(3.34)	(0.87)	(−1.18)	(10.74)	(13.16)	(1.96)	(2.58)	(0.87)
Decile 5	−0.120	1.014	1.284	0.126	0.049	0.91	−0.049	0.994	1.364	0.148	0.044	0.91
	(−0.27)	(15.80)	(13.68)	(1.73)	(3.01)	(0.91)	(−0.11)	(14.10)	(14.41)	(1.97)	(1.75)	(0.91)
Decile 6	0.287	1.005	1.322	0.054	0.011	0.95	0.270	0.985	1.346	0.047	0.026	0.95
	(0.90)	(20.90)	(19.68)	(1.03)	(0.97)	(0.95)	(0.85)	(20.33)	(20.66)	(0.91)	(1.46)	(0.95)
Decile 7	0.604	1.081	1.063	0.033	−0.004	0.97	0.588	1.077	1.059	0.027	0.001	0.97
	(2.57)	(32.08)	(21.57)	(0.86)	(−0.43)	(0.97)	(2.50)	(30.05)	(21.98)	(0.72)	(0.11)	(0.97)
Decile 8	0.843	1.063	0.840	−0.003	0.000	0.98	0.842	1.063	0.840	−0.003	0.000	0.98
	(4.27)	(37.60)	(20.32)	(−0.08)	(−0.03)	(0.98)	(4.27)	(35.36)	(20.80)	(−0.09)	(−0.00)	(0.98)
Decile 9	1.037	1.020	0.555	−0.023	−0.004	0.97	1.038	1.024	0.548	−0.022	−0.007	0.97
	(5.54)	(37.97)	(14.14)	(−0.75)	(−0.59)	(0.97)	(5.55)	(35.89)	(14.29)	(−0.74)	(−0.64)	(0.97)
Decile 10 (Large)	1.184	0.985	−0.145	−0.108	−0.001	0.98	1.178	0.983	−0.146	−0.110	0.001	0.98
	(9.64)	(55.93)	(−5.63)	(−5.40)	(−0.22)	(0.98)	(9.59)	(52.50)	(−5.80)	(−5.51)	(0.21)	(0.98)

Therefore, we conclude that our *PDI* factor is not fully spanned by other factors and has a sizable explanatory power in our regressions.

As for the economic significance of R_{PDI}, when R_{PDI} (constructed with value-weighted book-to-market portfolios) changes by one standard deviation, the excess return for the smallest non-financial firms changes by about 3.1 percent, versus 4.1 percent for the impact of market excess return. When we use R_{PDI} constructed with equal-weighted book-to-market portfolios, this number becomes 5.6 percent, which is larger than the impact of market excess return!

The magnitude of the coefficients on R_{PDI} in Table 14.8 is about the same as in Table 14.7, and this shows that without *SML* or *HML* in the regression, the *PDI* factor does not pick up higher loadings. This confirms that *PDI* risk factor represents an independent source risk which cannot be spanned by *SML* or *HML*.

14.6. CONCLUSION

An important message of Green and Porter (1984) is that collusion can be very subtle. The subsequent theoretical work is very elegant and powerful. See Abreu, Pearce, and Stacchetti (1990) and Fudenberg, Levine, and Maskin (1994). Empirical work on testing models of repeated games, however, has been difficult because of the data requirements for estimation of structural models. Empirical work has been limited and has focused on price wars as the only examples of such imperfect competition. We presented a theoretical model of strategic repeated bank lending, in which banks compete in a rather special way, via the intensity of information production about potential borrowers. Based on prior information, e.g., about bank loan interest rates being sticky, we conjectured which equilibrium occurred in reality. We then empirically tested the model by parameterizing the information on which banks' beliefs are based. The Performance Difference Indices are proxies for banks' beliefs.

We studied banking, an industry in which there have not been price wars. Banking is an industry with limited entry; it is a highly concentrated industry, and it is an industry that is informationally opaque and hence regulated. Banks produce private information about their borrowers, but they do not know how much information rival banks are producing. The information opaqueness affects competition for borrowers in that rivals can produce information with different precision. This causes the imperfect competition in banking to take a different form from other industries. In particular, we showed that the intertemporal incentive constraints implementing the collusive arrangement (of high interest rates and low cost information production) require periodic credit crunches.

Because banking is regulated, bank regulators collect information from banks, and release it at periodic intervals. So, information about rival banks is made

available by the government. All banks can see the performance of other banks. Our empirical approach to testing proceeds at the level of this public information that is the basis for banks' beliefs, changes in which cause credit cycles. Empirically we showed that a simple parameterization of relative bank performance differences has predictive power for rival banks behavior in the credit card market. Moreover, introducing the performance difference histories into a vector autoregression-type macroeconomic model, using commercial and industrial loans, confirms that this is an autonomous source of macroeconomic fluctuations.

Finally, since changes in bank beliefs based on public information cause credit cycles, this should be an important independent risk factor for stock returns, not only for banks but for borrowers. In an asset-pricing context this risk should be priced, even though it is endogenous. We showed that this is indeed the case. Smaller firms are more sensitive to this risk, confirming that such firms are more bank-dependent.

As mentioned in the Introduction, one topic for future research is the effects of monetary policy on the repeated bank lending game. Another topic is to find and analyze other instances where the same empirical strategy can be applied.

APPENDIX A-E: DETAILS OF THE REPEATED LENDING GAME AND PROOFS

A. Formalization of the Stage Strategy

Bank i randomly chooses n_i applicants to test. For those applicants that bank i does not test, it will decide to approve applications to $N_{ai} \leq N - n_i$ of the applicants, and offer the approved applicants a loan at interest rate F_{ai}. The bank rejects the rest of the non-tested applicants. For those applicants that are tested by bank i, the bank will observe a number of good type applicants, $N_{gi} \leq n_i$, and will then decide to approve applications to $N_{\beta i} \leq N_{gi}$ of the applicants that passed the test, and offer the approved applicants a loan at interest rate $F_{\beta i}$. Bank i can also decide to approve applications to $N_{\gamma i} \leq n_i - N_{gi}$ of the applicants that failed the test, and offer these approved applicants a loan at interest rate $F_{\gamma i}$. The bank rejects the remaining applicants. In general, $F_{\alpha i}$, $F_{\beta i}$ and $F_{\gamma i}$ could vary among the corresponding category of applicants, that is, different applicants in the same category could possibly get offers of loans at different interest rates. Therefore, we interpret $F_{\alpha i}$, $F_{\beta i}$, and $F_{\gamma i}$ as vectors of interest rates charged to those approved non-tested applicants. The stage strategy of a bank is:

$$s_i = \{n_i, N_\alpha(n_i, N_{gi}), N_\beta(n_i, N_{gi}), N_{\gamma i}(n_i, N_{gi}), F_{\alpha i}(n_i, N_{gi}), F_{\beta i}(n_i, \mathrm{N}_{gi}), F_{\gamma i}(n_i, N_{gi})\},$$

where:

n_i: the number of applicants that bank i tests;
N_{gi}: the number of good applicants found by bank i with the test;
$N_{\alpha i}$: the number of applicants that bank i offers loans to without test;
$N_{\beta i}$: the number of applicants that pass the test and get a loan from bank i;
$N_{\gamma i}$: the number of applicants that fail the test and get a loan from bank i;
$F_{\alpha i}$: the interest rate on the loan that bank i offers to the applicants without a test;
$F_{\beta i}$: the interest rate on the loan that bank i offers to the applicants that pass the test;
$F_{\gamma i}$: the interest rate on the loan that bank i offers to the applicants that fail the test.

B. Proof of Proposition 1

We first prove the following lemma.

LEMMA 1 *If it exists, in any symmetric stage Nash equilibrium in which neither bank conducts credit worthiness testing, each bank offers loans to all the loan applicants at the same interest rate.*

Proof. It is easy to check that if bank i is playing $s_i = (n_i = 0, N_{\alpha i} < N, F_{\alpha i})$, then bank $-i$ can strictly increase its profits by playing $s'_{-i} = (n_{-i} = 0, N'_{\alpha -i} = N',\ F'_{\alpha -i})$, where the strategy is $s'-_i$ to offer $F'_{\alpha -i} = F_{\alpha i}$ to $N_{\alpha i}$ applicants (although these $N_{\alpha i}$ applicants might not be the same applicants that bank i is offering loans to), and offer X to the rest of them. Let F^* be the interest rate corresponding to zero profits in the loan market when there is no testing. Then:

$$E\pi_i = \frac{N}{2}[\lambda p_b F^* + (1-\lambda)p_g F^* - 1] = 0,$$

$$\text{and } F^* = \frac{1}{\lambda p_b + (1-\lambda)p_g} < X (\text{by Assumption 1}).$$

Assume bank i is playing $s_i = (n_i = 0, N_{\alpha i} < N, F_{\alpha i})$, with $F_{\alpha i} = (F_1, F_2, \ldots, F_N)$. Suppose $F_j \geq F^*$ for $j = 1, 2, \ldots, N$ and assume there exist j and k, such that $F_j \neq F_k$, and, without loss of generality, $F_k \geq F^*$. Bank $-i$ can strictly increase its profitability by playing $s'_{-i} = \left(n_{-i} = 0, N'_{\alpha -i} = N, F'_{\alpha -i}\right)$, where $F_{\alpha i} = \left(F_1, \ldots, F_{k-1}, F_k^-, F_{k+1}, \ldots, F_N\right)$ and $F_k{}^-$ is smaller than F_k by an infinitely small amount. Therefore, interest rates are bid down until each bank offers F^* to all the applicants.

PROOF PROPOSITION 1: From Lemma 1, we see that in a symmetric equilibrium with no bank testing applicants, both banks offer loans to all the applicants

at $F^* = \frac{1}{\lambda p_b + (1-\lambda) p_g} < X$ (by Assumption 1). With $c < \frac{(1-\lambda)\lambda\left(p_g - p_b\right)}{\lambda p_b + (1-\lambda) p_g}$, a bank will have an incentive to conduct credit worthiness testing on at least one loan applicant and to offer loans to those applicants that pass the test, offering an interest rate F^{*-}, which is lower than F^* by an infinitely small amount. To see this consider a bank that deviates by conducting credit worthiness testing on one applicant. The expected profit from this deviation is:

$$E\pi_i^d = (1-\lambda)\left(p_g F^* - 1\right) - c.$$

We have:

$$E\pi_i^d > 0 \,\text{iff}\, c < (1-\lambda)\left(p_g F^* - 1\right) = \frac{(1-\lambda)\,\lambda\left(p_g - p_b\right)}{\lambda p_b + (1-\lambda) p_g}.$$

We can see that if $c \geq \frac{(1-\lambda)\lambda\left(p_g - p_b\right)}{\lambda p_b + (1-\lambda) p_g}$, then F^* will be a Nash equilibrium interest rate on the loan, and no bank will conduct credit worthiness testing.

C. Proof of Proposition 2

We first prove the following three lemmas.

LEMMA 2 *In any symmetric stage Nash equilibrium in which both banks test all the applicants, each bank offers loans to all the applicants that pass the test at the same interest rate.*

The proof is similar to **Lemma 1** and is omitted.

LEMMA 3 *If it exists, in any symmetric stage Nash equilibrium in which both banks test $n<N$ applicants, each bank offers loans to all applicants that pass the test (good types) at $F^{**} = \frac{1}{p_g}$.*

The proof is similar to **Lemma 1** and is omitted.

LEMMA 4 *If it exits, in any symmetric stage Nash equilibrium in which both banks test $n < N$ applicants, each bank either offers loans to all non-tested applicants at the same interest rate or offers loans to none of them.*

Proof. If there exists a feasible $F \leq X$ such that the banks can make a strictly positive profit by lending to non-tested applicants at F, following a similar argument as in the proof of Lemma 1, we conclude that each bank offers loans to all non-tested applicants at the same interest rate. If there does not exist a feasible F such that the banks can make a non-negative profit by lending to non-tested applicants at F, we conclude that each bank offers loans to none of those non-tested applicants.[30]

30. Here we neglect a non-generic case in which there exists an F such that the banks can earn zero profit by offering loans to a non-tested applicant, and there does NOT exist an F such that the banks can earn strictly positive profit by offering loans to a non-tested applicant. In this case, each

PROOF PROPOSITION 2: The proof is by contradiction. If in equilibrium both banks conducting credit worthiness testing on all the applicants, from Lemma 2, both banks offer loans to all the applicants that pass the test, i.e., $N_\beta = N_g$, where N_g denotes the number of applicants passing the test. Banks will make no loans to bad types found by testing, that is, $N_\gamma = 0$. Both banks use the credit worthiness test at a cost c per applicant. Assume the loan interest rate they charge to approved applicants is $F_\beta\left(N, N_g\right)$, depending on N_g. Each bank must earn non-negative expected profits $E\pi \geq 0$, i.e., the participation constraints. For each realization of N_g, each bank expects to make loans to $N_g/2$ applicants. Let p_k denote the probability of finding k good type applicants. Then:

$$E\pi_i = E\sum_{k=0}^{N} \frac{1}{2} k p_k \left[p_g F_\beta\left(N,k\right) - 1\right] - Nc \geq 0.$$

Assume now, if bank i cuts F_β by an infinitely small amount, that is, $F_\beta^d\left(N_g\right) = F_\beta^-\left(N_g\right)$, then it will loan to N_g applicants for any realization of N_g. We have:

$$E\pi_i^d = E\sum_{k=0}^{N} k p_k \left[p_g F_\beta^-\left(N,k\right) - 1\right] - Nc \geq E\pi_i.$$

For the case in which both banks conducting credit worthiness testing on a subset of the applicants, if the banks offer loans all non-tested applicants, we have $F_\beta = F^{**}$ and $F_\alpha = F\left(n\right)$, which are the interest rate that results in zero expected profit from offering loans to tested good type applicants and non-tested applicants when banks test n applicants. It is easy to check that $F\left(n\right) > F^{**}$. The argument for $F_\alpha = F\left(n\right)$ is similar to the argument for $F_\beta = F^{**}$. However, at $F_\alpha = F\left(n\right)$ and $F_\beta = F^{**}$, banks will earn negative expected profit due to the test cost. If the banks offer loans to none of the non-tested applicants, the banks will only offer loans to those applicants that passed the test at F^{**}. The argument is similar.

D. Formalization of the Repeated Game

Assume that the two banks play the lending market stage game period after period, each with the objective of maximizing its expected discounted stream of profits. Upon entering a period of play, a bank observes only the history of:

(i) its own use of the credit worthiness test and the results;

bank can possibly offer to a subset of the non-tested applicants. However, including this case will not affect the results in Proposition 1.

(*ii*) its own interest rate on the loan offered to applicants;
(*iii*) its own choice of applicants that it lent to;
(*iv*) its own and its competitor's loan portfolio size (number of loans made);
(*v*) its own and its competitor's number of successful loans.

For bank *i,* a full path play is an infinite sequence of stage strategies. The infinite sequence $\{s_{it}\}_{t=0}^{\infty}, i = 1,2$, together with nature's realization of the number of good type applicants and the applicants' rational choice of bank, implies a realized sequence of loans from bank *i,* as well as a quality of the borrowers who received loans from bank *i*. That is:

$$K_{it} = (D_{\alpha it}, D_{\beta it}, D_{\gamma it}, \chi_{\alpha it}, \chi_{\beta it}, \chi_{\gamma it}),$$

where D denotes the number of applicants that accepted the offer, and χ denotes the number of successful borrowers; α, β, and γ denote the corresponding category, as defined earlier ($\alpha \equiv$ untested, approved, applicants; $\beta \equiv$ tested, good types, approved; $\gamma \equiv$ tested, bad types, approved). Define:

$$D_{it} = D_{\alpha it} + D_{\beta it} + D_{\gamma it}$$
$$\chi_{it} = \chi_{\alpha it} + \chi_{\beta it} + \chi_{\gamma it}.$$

Let the public information at the start of period $t + 1$, be $\kappa_t = (\kappa_{1t}, \kappa_{2t})$, where $\kappa_{it} = (D_{it}, \chi_{it}), i = 1,2$ (for each bank). So, the information set includes the realization of the number of loans made by bank *i* and the number of borrowers that repaid their loans in period t.

At the beginning of period T bank i has an information set: $h_i^{T-1} = \left\{\alpha_{it}, K_{it,}\kappa_t\right\}_{t=0}^{T-1} \in H_i^{T-1}$, where $a_{it} = \left\{n_{it}, N_{\alpha it}, N_{\beta it}, N_{\gamma it}, F_{\alpha it}, F_{\beta it}, F_{\gamma it}\right\}$ is the action of bank i (by convention $h_i^{-1} = \phi$). A (pure) strategy for bank i associates a schedule $\sigma_{iT}\left(h_i^{T-1}\right)$ with each T = 0,1, . . . and $\sigma_{iT} : H_i^{T-1} \to S$, where S is the stage strategy space with element s_{it}, defined earlier. Denote the public information as $h^{T-1} = \{\kappa_t\}_{t=0}^{T-1} \in H^{T-1}$, and a (pure) strategy for bank i associates a schedule $\sigma_{iT}\left(h^{T-1}\right)$ with each $T = 0, 1, \ldots$ and $\sigma_{iT} : H_i^{T-1} \to S$.

Given λ, p_g, and p_b (that is, nature's uncertainty), a strategy profile (σ_1, σ_2), with $\sigma_i = \{\sigma_{it}(.)\}_{t=0}^{\infty}, i = 1,2$, recursively determines a stochastic process of credit standards $\left(\{n_{it}\}_{t=0}^{\infty}, i = 1,2\right)$, interest rates $\left(\{F_{it}\}_{t=0}^{\infty}, i = 1,2\right)$, bank portfolio sizes and loan outcomes $\left(\{\kappa_{it}\}_{t=0}^{\infty}, i = 1,2\right)$. The expected pathwise payoff for bank i is:

$$v_i(\sigma_1, \sigma_2) = E\sum_{i=0}^{\infty} \delta^t \pi_i(s_{1t}, s_{2t}),$$

where

$$\pi_i(s_{1t}, s_{2t}) = (\chi_{\alpha it} F_{it} - D_{\alpha it}) + \left(\chi_{\beta it} F_{it} - D_{\beta it}\right) + \left(\chi_{\gamma it} F_{it} - D_{\gamma it}\right) - n_{it} c.$$

E. Definition of Symmetric Perfect Public Equilibrium

A Perfect Public Equilibrium (PPE) is a profile of public strategies that, starting at any date t and given any public history h_i^{t-1}, forms a Nash equilibrium from that point on (see Fudenberg, Levine, and Maskin (1994)).

As shown by Abreu, Pearce, and Stacchetti (1990), any perfect public equilibrium payoff for bank i can be factored into a first-period stage payoff π_i(depending on the stage strategies of both banks) and a continuation payoff function u_i (depending on the public history). Let s_i be the stage strategy for bank i, a symmetric perfect public equilibrium (SPPE) is defined as follows:

DEFINITION: *A Symmetric Perfect Public Equilibrium (SPPE) is a Perfect Public Equilibrium that can be decomposed into the first period stage strategies and continuation value functions* (s_1, s_2, u_1, u_2) *such that:*

$$s_1 = s_2 \text{ and } u_1(D_1, D_2, \chi_1, \chi_2) = u_2(D_2, D_1, \chi_2, \chi_1).$$

According to the definition, the stage game strategies are the same, but the continuation strategies can differ. In particular, note that the continuation value functions for Bank 1 and Bank 2 are symmetric in that if we exchange the loan portfolio sizes and loan performances, the continuation values will also be exchanged. In such an SPPE, the expected payoff for the two banks are the same, but asymmetric play is allowed after the first period, for asymmetric realizations of loan portfolio size and loan performance.

LEMMA 5 *In a Symmetric Perfect Public Equilibrium, if on the equilibrium path, banks make offers to all loan applicants without credit worthiness tests at an interest rate higher than* $F^* = \frac{1}{\lambda p_b + (1-\lambda) p_g}$, *and the continuation payoffs only depend on loan portfolio distribution* (D_1, D_2), *then for any value of D we have:*

$$\delta u_i(D, N-D) - \delta u_i(D+1, N-D-1) = \left[\lambda p_b + (1-\lambda) p_g\right] F_\alpha - 1.$$

Proof: Assume that there exists a SPPE with $s = (n = 0, N_\alpha = N, F_\alpha)$ played on the equilibrium path, where F_α is a constant larger than $F^* = \frac{1}{\lambda p_b + (1-\lambda) p_g}$, and the continuation value function does not depend on (χ_1, χ_2), which are the numbers of defaulted loans in banks' loan portfolios. To eliminate the incentive for a bank i to deviate to strategy $s'(D) = \left(n = 0, N_\alpha = N, F_\alpha^-\right)$ with $0 \le D \le N$, for any $D \ne D'$, we must have:

$$\pi_i\left(s'(D), s\right) + \delta u_i(D, N-D) = \pi_i\left(s'\left(D'\right), s\right) + \delta u_i\left(D', N-D'\right),$$

which implies:

$$\delta u_i(D, N-D) - \delta u_i(D+1, N-D-1) = \pi_i\left(s'(D+1), s\right) - \pi_i\left(s'\left(D'\right), s\right).$$

The result is immediate. Intuitively, the expected payoff with no deviation is a linear combination of the expected payoffs with deviations in the form of $s'(D)$, $D = 0, 1, \ldots, N$. Therefore, the expected payoff for each deviation with $s'(D)$ must be the same.

APPENDIX F: DETAILS OF THE BOOTSTRAP

For each round of the bootstrap, the Significance Index is constructed as follows. For each of the 30 pairwise regressions, when the average coefficient of Z_{ijt} is negative, if the chi-squared-statistic is significant at the 99% confidence level, add a value of 4 to *SI,* if it is significant at the 95% confidence level, add a value of 3 to *SI,* if it is only significant at the 90% confidence level, add a value of 2 to *SI,* and add a value of 1 otherwise; when the average coefficient of Z_{ijt} is negative, if the chi-squared-statistic is significant at the 99% confidence level, add a value of –4 to *SI*, if it is significant at the 95% confidence level, add a value of –3 to *SI*, if it is only significant at the 90% confidence level, add a value of –2 to *SI,* and add a value of –1 otherwise.[31] The index SI takes care of both the significance and the sign of the coefficients of z_{ijt}. If the *p*-value of SI^* is small enough, we reject the Null hypothesis and accept the alternative one.

The bootstrap algorithm is as follows:

Step 1: Run the OLS regression in H_0, for the two cases where $y_{it} = LL_{it}$ or LR_{it}, and use the estimated coefficients, α_{OLS}, to generate the residuals ${u_{it}}^*$.

Step 2: We can sample from ${u^*}_{it}$ in the regressions to generate new LL^*_{it} or LR^*_{it} using $y^*_{it} = \alpha_i x^*_{it} + u^*_{it}$ This also creates new ${x_{it}}^*$ and ${z_{ijt}}^*$ since both variables involve lags of LL_{it} and LL_{jt}.

Step 3: Use ${y^*}_{it}$, ${x_{it}}^*$, and ${z^*}_{it}$ from bootstrap to run the pairwise regression in H_1, and calculate the Significant Index SI.

Step 4: Repeat Step 2 to Step 3 100,000 times, and obtain the distribution of SI.

Step 5: Calculate the *p*-value of SI^*, i.e. $\Pr(SI = SI^*)$.

REFERENCES

Abreu, Dilip, David Pearce, and Ennio Stacchetti (1990), "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring," *Econometrica* 58 (5), 1041–63.

31. Admittedly there is some arbitrariness in how the Significance Index is constructed. However, we tried constructing the Significance Index in a number of ways, and found that the results are robust.

Asea, Patrick K., and S. Brock Blomberg (1998), "Lending Cycles," *Journal of Econometrics* 83: 89–128.

Asgharian, Hossein (2006), "A Comparative Analysis of the Ability of Mimicking Portfolios in Representing the Background Factors," Working Paper, Lund University.

Ashcraft, Adam (2003), "New Evidence on the Lending Channel," Federal Reserve Bank of New York, working paper.

Ausubel, Lawrence M. (1991), "The Failure of Competition in the Credit Card Market," *American Economic Review* 81 (1), 50–81.

Beatty, Anne L and Anne Gron (2001), "Capital, Portfolio, and Growth: Bank Behavior Under Risk-Based Capital Guidelines," *Journal of Financial Services Research* 20 (1), 5–31.

Beck, Thorsten, Asli Demirguc-kunt, and Ross Levine (2003), "Bank Concentration and Crises," World Bank Working Paper.

Berger, Allen N. and Gregory F. Udell (1992), "Some Evidence on the Empirical Significance of Credit Rationing," *Journal of Political Economy* 100 (5), 1047–1077.

Berger, Allen N. and Gregory F. Udell (1994), "Do Risk-Based Capital Allocate Bank Credit and Cause a 'Credit Crunch' in the United States?" *Journal of Money, Credit and Banking* 26 (3), 585–628.

Bernanke, Ben S. and Alan Blinder (1988), "Credit, Money, and Aggregate Demand," *American Economic Review* 78, 435–39.

Bernanke, Ben S. and Mark Gertler (1995), "Inside the Black Box: The Credit Channel of Monetary Policy," *Journal of Economic Perspectives* 9(4), 27–48.

Bernanke, Ben S. and Cara S. Lown (1991), "The Credit Crunch," *Brookings Papers on Economic Activity* 2, 204–39.

Breeden, Douglas T., Michael R Gibbons, and Robert H. Litzenberger (1989), "Empirical Tests of the Consumption-Oriented CAPM," *Journal of Finance* 44, 231–62.

Bresnahan, Timothy (1989), "Empirical Studies of Industries with Market Power," in R. Schmalensee and R.D. Willing, eds., *Handbook of Industrial Organization,* Vol. 2, p. 1011–57 (New York: North Holland).

Brinkmann, Emile J. and Paul M. Horvitz (1995), "Risk-Based Capital Standards and the Credit Crunch," *Journal of Money, Credit & Banking* 27 (3), 848–63.

Broecker, Thorsten (1990), "Credit-Worthiness Tests and Interbank Competition," *Econometrica* 58 (2), 429–52.

Carhart, Mark (1997), "Persistence in Mutual Fund Performance," *Journal of Finance* 52, 57–82.

Cochrane, John (1999), "New Facts in Finance," *Economic Perspectives* XXIII (3) (Federal Reserve Bank of Chicago).

Cronshaw, Mark and David G. Luenberger (1994), "Strongly Symmetric Subgame Perfect Equilibrium in Infinitely Repeated Games with Perfect Monitoring and Discounting," *Games and Economic Behavior* 6, 220–37.

Dell'Ariccia, Giovanni and Robert Marquez (2004), "Lending Booms and Lending Standards," *Journal of Finance,* forthcoming.

Fama, Eugene and Kenneth French (1993), "Common Risk Factors in the Returns of Stocks and Bonds," *Journal of Financial Economics* 33, 3–56.

Fama, Eugene and Kenneth French (1996), "Multifactor Explanations for Asset Pricing Anomalies," *Journal of Finance* 51, 55–94.

Fudenberg, Drew, David I. Levine, and Eric Maskin (1994), "The Folk Theorem with Imperfect Public Information," *Econometrica*, 62 (5), 997–1039.

Furfine, Craig (2001), "Bank Portfolio Allocation: The Impact of Capital Requirements, Regulatory Monitoring, and Economic Conditions," *Journal of Financial Services Research* 20 (1), 33–56.

Gorton, Gary B. and Andrew Winton (2003), "Financial Intermediation," in *The Handbook of the Economics of Finance:* Corporate Finance, edited by George Constantinides, Milton Harris, and Rene Stulz (Elsevier Science; 2003) (NBER Working Paper # 8928).

Green, Edward J. and Robert H. Porter (1984), "Noncooperative Collusion under Imperfect Price Information," *Econometrica* 52 (1), 87–100.

Gross, David and Nicholas Souleles (2002), "An Empirical Analysis of Personal Bankruptcy and Delinquency," *Review of Financial Studies* 15, 319–47.

Group of Ten, "Report on Consolidation in the Financial Sector," January 25, 2001.

Hall, Brian J. (1993), "How Has the Basle Accord Affected Bank Portfolios?" *Journal of the Japanese and International Economics* 7, 408–40.

Hamilton, James D. (1994), "Time Series Analysis," Princeton University Express.

Hancock, Diana and James A. Wilcox (1994), "Bank Capital and the Credit Crunch: The Roles of Risk-Weighted and Unweighted Capital Regulations," *Journal of the American Real Estate & Urban Economics Association,* 22 (1), 59–94.

Hancock, Diana and James A. Wilcox (1998), "The "Credit Crunch" and the Availability of Credit to Small Business," *Journal of Banking and Finance* 22, 983–1014.

Haubrich, Joseph and Paul Wachtel (1993), "Capital Requirements and Shifts in Commercial Bank Portfolios," *Economic Review* (Federal Reserve Bank of Cleveland), 29, 2–15.

Horowitz, Joel L. (2001), "The Bootstrap," *Handbook of Econometrics,* Vol. 5, J.J. Heckman and D. E. Leamer, eds., Elsevier Science B.V., Ch. 52, 3159–3228.

Jordan, John, Joe Peek, and Eric Rosengren (2002), "Credit Risk Modeling and the Cyclicality of Capital," Federal Reserve Bank of Boston, working paper.

Kreps, David M. and Robert Wilson (1982), "Sequential Equilibria," *Econometrica,* 50 (4), 863–94.

Keeton, William R. (1994), "Causes of the Recent Increase in Bank Security Holdings," *Economic Review* (Federal Reserve Bank of Kansas City), 79 (2), 45–57.

Lown, Cara and Donald P. Morgan (2005), "The Credit Cycle and the Business Cycles: New Findings Using the Survey of Senior Loan Officers," *Journal of Money, Credit & Banking,* forthcoming.

Lown, Cara and Donald Morgan (2002), "Credit Effects in the Monetary Mechanism," *Economic Policy Review,* 8(1) (May 2002), Federal Reserve Bank of New York.

Lown, Cara, Donald Morgan, and Sonali Rohatgi (2000), "Listening to Loan Officers: The Impact of Commercial Credit Standards on Lending and Output," *Economic Policy Review*, 6 (July 2000), 1–16.

Peek, Joe and Eric Rosengren (1995), "The Capital Crunch: Neither an Applicant Nor a Lender Be," *Journal of Money, Credit & Banking,* 27 (3), 625–38.

Rajan, Raghuram G. (1994), "Why Bank Credit Policies Fluctuate: A Theory and Some Evidence," *The Quarterly Journal of Economics,* 109 (2), 399–441.

Reiss, Peter and Frank Wolak (2005), "Structural Econometric Modeling," *Handbook of Econometrics*, Volume 6, forthcoming.

Ruckes, Martin (2004), "Bank Competition and Credit Standards," *Review of Financial Studies,* 17, 1073–1102.

Schreft, Stacey L. and Raymond E. Owens (1991), "Survey Evidence of Tighter Credit Conditions: What Does It Mean?" *Federal Reserve Bank of Richmond Economic Review,* 77 (2), 29–34.

Weinberg, John A. (1995), "Cycles in Lending Standards?" *Federal Reserve Bank of Richmond Economic Quarterly*, 81 (3), 1–18.

PART IV

Change in Banking

15

Corporate Control, Portfolio Choice, and the Decline of Banking

GARY B. GORTON AND RICHARD ROSEN*

The 1980s was not a good decade for U.S. banks. Gerald Corrigan (1992), the head of the New York Federal Reserve Bank during the period, observed that: "... we would all accept the fact that the decade of the 1980s was surely the most difficult interval faced by the U.S. banking system since the 1930s." Indeed, during the 1980s, bank profitability declined steadily, whether measured by accounting return on equity, return on assets, or market value. Figure 15.1 shows the accounting return on assets.[1] Not only did banking

* Gorton is from The Wharton School, University of Pennsylvania and the National Bureau of Economic Research (NBER). Rosen is from Indiana University. Thanks to Stephen Buser, Charles Calomiris, Frank Diebold, Mark Flannery, Javier Hidalgo, Chris James, Myron Kwast, David Llewellyn, Max Maksimovic, Pat McAllister, George Pennacchi, Steve Prowse, Rene Stulz, Greg Udell, an anonymous referee, and participants of seminars at the London School of Economics, Stockholm School of Economics, the Board of Governors Lunchtime Workshop, the Penn Macro Lunch Group, the University of Chicago, the Chicago Fed Bank Structure Conference, Cornell University, University of Florida, University of Michigan, the NBER Corporate Finance Group, the Maryland Symposium, the Office of Thrift Supervision, and the San Francisco Federal Reserve Bank for suggestions and discussion. Much of the work on this paper was done while Rosen was at the Board of Governors of the Federal Reserve System. The views expressed in this paper represent the authors' views only and do not necessarily represent the views of the Federal Reserve System. Part of this paper was previously part of a paper entitled "Overcapacity and Exit From Banking."

1. Controlling for the effects in 1987 and 1988 of large bank write-downs of LDC loans in 1987, the decline in profits shown in Figure 15.1 is statistically significant. The increase in charge-offs is also significant. Market value data on the return to bank equity is consistent with the book value data shown in Figure 15.1. Over the 1980s the S&P 500 outperformed the Salomon Brothers index

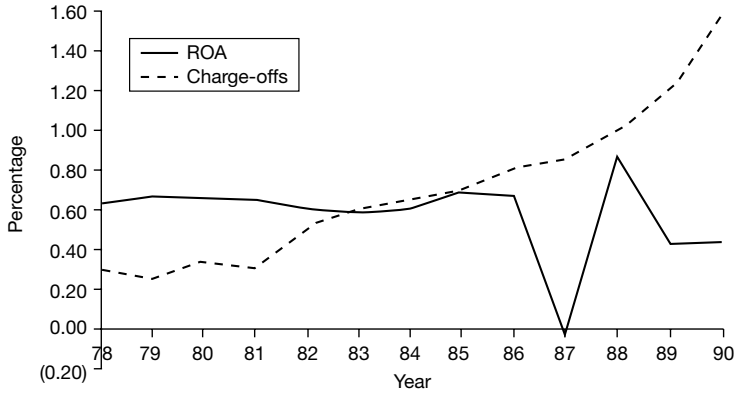


Figure 15.1 Bank Return on Assets and Bank Charge-Offs (for banks larger than $300 million in assets, 1978–1990). The data source is the *Call Reports of Income and Condition.*

become less profitable, it became riskier. The ratio of charge-offs to total loans, a measure of risk, rose almost monotonically in the last decade. (See Figure 15.1.)

Not surprisingly, bank failures, which averaged six (mostly small banks) per year from 1946 to 1980, rose exponentially, averaging 104 banks per year during the 1980s.[2] Unlike the 1930s, however, it is not obvious what caused the recent decade of malaise in the industry.

The 1990s have seen a turnaround in bank prospects. But the increased profits appear to largely be due to short-term phenomena that may not affect the long-term decline in banking.[3] Declining interest rates in 1991 and 1992 allowed banks to profit from the sale of investment securities. See *Federal Reserve Bulletin* (July 1993). Interest margins also increased during the same period. If interest rates rise from their current levels, banking may return to 1980s profit levels. As Corrigan (1992) observes, rebuilding the U.S. banking system is likely to be a long and difficult process.

Our concern is with the low bank profits during the 1980s (we discuss the 1990s further in the conclusion). The decline of U.S. bank profitability in the 1980s coincided with significant changes in corporate finance. Banks, in particular, lost market share in financing corporations, one of their core lending areas.

of bank stocks by 38 percent. Also, see Table 15.2, discussed later in the text, for data on the return on loans.

2. Bank failure figures are from the FDIC *Annual Report*. FDIC payouts show a similar trend.

3. It is important to be clear about what we mean by "banking" being in decline. The term "banking" has traditionally corresponded to a particular set of activities, namely, financing loans by issuing deposits. The combination of these activities has, historically, been the source of public policymakers' concerns. As we discuss, there has been a decline in corporate lending by banks and, because of money market mutual funds, a smaller decline in demand deposits.

In the past, banks had been the dominant providers of short-term (nonfinancial) corporate debt. But their share of this market has been declining, from about 70 percent in the late 1970s to less than 60 percent by the late 1980s. Theoretical work suggests that bank loans are the most efficient method of supplying capital in the presence of information or monitoring problems.[4] Historically, corporations have been prone to these sorts of problems. Technological change or changes in market structures may have reduced the information and monitoring problems for many corporations, meaning there is less need for bank loans to finance these borrowers.[5] These changes have allowed many large and medium-sized firms to access nonbank capital markets.[6]

Banks should have responded to the changes in the corporate debt market by reducing the volume of corporate loans while seeking new profit opportunities to replace lost opportunities. In fact, there has been a shift in bank portfolios, to off-balance sheet activity, such as loan commitments and standby letters of credit for corporations.[7] Banks also significantly increased commercial real estate lending in recent years. Commercial real estate more than doubled, as a percentage of total bank assets, between 1980 (when the percentage was 5.36) and 1990 (when it was 11.13). But, these changes were not enough to replace lost bank profit.

Why did banking become unprofitable, and bank failures increase, in the 1980s? A large literature in banking, following Merton (1977), concentrates on the incentives of shareholders to maximize the value of the (fixed rate) deposit insurance subsidy provided by the government by taking on risk inefficiently, so-called "moral hazard" risk.[8] As refined by Marcus (1984) and Keeley (1990), bank shareholders have an incentive to take on risk when the value of the bank

4. Theoretical work on banking argues that commercial banks can produce information about potential borrowers and monitor the managements of borrowing firms, by enforcing loan covenants, in ways which cannot easily be replicated by marketable, corporate securities. See Boyd and Prescott (1986) and Diamond (1984). Bhattacharya and Thakor (1993) provide a review. The empirical evidence that bank loans are unique includes James (1987) and Lummer and McConnell (1989). Also, see Hoshi, Kashyap, and Scharfstein (1990), Gilson, John, and Lang (1990), James and Weir (1991), and Fama (1985).

5. Gorton and Pennacchi (1990), studying the loan sales market, provide some evidence for this proposition.

6. However, small firms and retail customers are relatively unaffected by the technological changes. Thus, banks that lend primarily to smaller firms, particularly small banks, might not be subject to many of the problems we discuss here.

7. Standby letters of credit, letters of credit, foreign exchange commitments, commitments to make loans, futures and forward contracts, options, and swaps, all show significant upward time trends over the 1980s. Some of these categories have increased dramatically.

8. It should be stressed that empirical research has not reached a consensus on whether deposit insurance is underpriced (see Marcus and Shaked (1984), Ronn and Verma (1986), and Pennacchi (1987)).

charter falls sufficiently (Keeley claims that charter values have fallen recently; this is consistent with the decline in bank profitability).

In this paper we take issue with the view that moral hazard emanating from fixed rate deposit insurance explains the recent behavior of the U.S. banking industry. The moral hazard view of banks assumes that shareholders make the lending decisions and can take on risk to maximize the value of insurance if they desire. Rather than assume that shareholders directly control bank actions, we assume bank managers, who may own a fraction of the bank, make the lending decisions. If managers have different objectives than outside shareholders and disciplining managers is costly, then managerial decisions may be at odds with the decisions outside shareholders would like them to take.[9] We explore the effect of this conflict on the risk-taking behavior of banks.

The agency relationship between managers and outside shareholders has been widely studied in corporate finance. Jensen and Meckling (1976) and others argue that managers benefit from control of the firm in many ways, including the ability to consume nonmarketable perquisites. To protect future private benefits, and because managers have a large undiversifiable stake in the firm that employs their human capital, managers of nonfinancial firms avoid risk. Private managerial benefits of control, however, can be mitigated if managers' objectives are aligned with the objectives of outside shareholders. One way in which alignment of interests may occur is through managerial ownership of the firm's stock.

The trade-off between private benefits and ownership rewards is complicated since stockholding by managers who are not majority owners may actually increase their ability to resist monitoring, rather than serve to align the interests of outside equity owners and managers. Several studies of nonfinancial firms predict (Stulz (1988)) or find a nonlinear relationship between insider ownership and firm value reflecting this trade-off. Morck, Shleifer, and Vishny (1988) examine the effect of insider concentration on nonfinancial firms, as measured by Tobin's q. They impose a piecewise linear relationship and find that as insider ownership rises up to 5%, q increases; then q falls as the insider concentration grows to 25 percent; finally, it again rises at higher ownership levels. They interpret these results as showing the balance of three factors. For small insider holdings, the incentives of insiders become more aligned with those of the outsiders, but management does not have enough power to be entrenched.

9. If a bank's (market-value) capital ratio is sufficiently low, then both managers and outside shareholders may agree that the bank should maximize the value of deposit insurance. We do not dispute this argument. Our focus is on the prior question of how the bank came to have a low capital ratio. Consequently, we study banks which satisfy regulatory capital requirements. For the banks we study, the interests of managers and outside shareholders may be in conflict and it is not obvious that outside shareholders are able to induce managers to increase risk at the expense of the government, even if they want to.

As insider concentration continues to rise, management becomes entrenched. Equity shares are large enough to stave off effective outside disciplining, but not so large that management interests are the same as those of outside shareholders. A further increase in concentration aligns management interests with outsiders; managers essentially become the sole owners.

McConnell and Servaes (1990), examining nonfinancial firms, impose a quadratic relationship between Tobin's q and the concentration of both insider and outsider holdings. They find that q initially rises, and then falls as interests between the inside managers and outside shareholders become aligned. Finally, Saunders, Strock, and Travlos (1990) estimate a linear relationship between insider ownership and portfolio choice for a sample of 38 bank holding companies. They find that "stockholder controlled" banks took on more risk than "managerially controlled" banks.[10]

The varying specifications of the relationship between insider stockholding and firm performance motivates the model and the empirical tests we develop in this paper. We propose a model of corporate control in banking which has the crucial feature that investment opportunities have deteriorated: there are relatively fewer "good" lending opportunities. This allows us to be precise about the source of value reduction, namely, the risk and return choices made by bank managers facing deteriorating investment opportunities.

The decline in investment opportunities means that for banks there are fewer positive net present value (NPV) loans to be made than previously. The presence (or absence) of positive NPV lending opportunities may be an attribute of individual banks which have retained profitable customers or of individual bank managers who have the ability to locate these opportunities. In reality it is probably a combination of these factors. For our purposes this distinction is not important, but in the model we assume an "unhealthy" banking industry is one with a large proportion of low quality ("bad") managers. We interpret this as reflecting these poor investment opportunities. (The model may be slightly reinterpreted as reflecting qualities of banks rather than managers, as discussed below.)

When investment opportunities are declining, managers behave differently than in "healthy" industries (see Jensen (1993)). This is particularly true in banking, where asymmetric information and deposit insurance mean that banks can continue to issue liabilities (i.e., insured demand deposits) even if there are few good lending opportunities. The risk-avoiding behavior of managers stressed in the corporate finance literature presumes that conservative behavior is sufficient for job and perquisite preservation. When bad managers

10. Also see Bagnani, Milonas, Saunders, and Travlos (1994) who study the interaction of managerial ownership and risk-taking by analyzing how managerial ownership and bond yields are related.

predominate, conservative behavior may not allow most managers to keep their jobs and perquisites. These managers may find it optimal to take excessively risky actions. Thus, aggregate risk-taking, driven by attempts by bad managers to convince shareholders that they are good managers, can be excessive (relative to a first-best world and, perhaps, relative to an unregulated industry).

Our model and empirical work analyzes conflicts between managers and shareholders of solvent banks. Note that when banks have low capital ratios both the managers and the shareholders want to take risky actions if deposit insurance offers a subsidy for risk-taking. This is the "moral hazard" that many argue existed in the thrift industry after capital ratios fell dramatically with increases in interest rates in the 1970s. We do not dispute the logic of this argument for commercial banks when capital ratios are low and deposit insurance is fixed price. The difficulty with this explanation for commercial bank performance, however, is that it does not explain how banks came to have low capital ratios. We study well-capitalized banks and argue that our model and empirical results can explain how many banks came to have low capital ratios in the 1980s.

Section 15.1 sets out the game between a bank manager and shareholders and solves for a sequential Nash equilibrium. Section 15.2 discusses the assumptions of the model. The model makes specific predictions about the types of loans that managers make as a function of how much stock they own in the bank and as a function of the risk and return characteristics of different loan types. In Section 15.3 we discuss how this allows us to distinguish empirically the corporate control hypothesis from the moral hazard hypothesis. Tests of the model are reported on in Section 15.4. Section 15.5 concludes.

15.1. A MODEL OF BANKING LENDING DECISIONS

In this section we discuss a model of bank lending in which managers, not outside shareholders, make lending decisions. The managers receive private benefits from control of the bank and it is costly for outside shareholders to fire them. The cost of firing faced by outside shareholders increases with the extent to which managers own stock in the bank.

15.1.1. The Lending Environment

There are three dates and many banks. Each bank is run by a manager who has $1 to invest. Investment opportunities in banking vary either because loan opportunities are locationally or specialty dependent or because managers have different abilities for locating various types of lending opportunities. We model the heterogeneity in opportunities as a function of manager type although we discuss

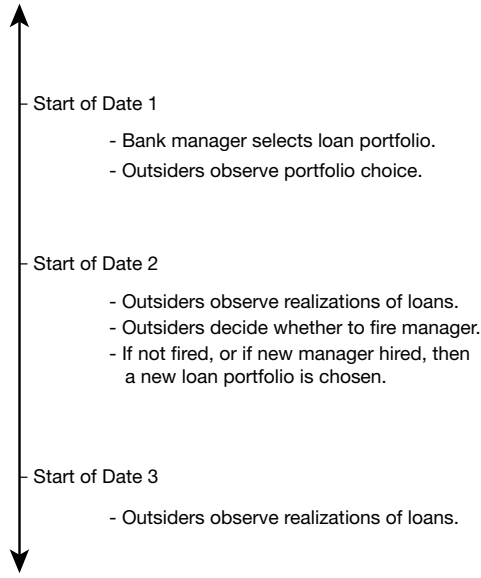


Figure 15.2 Sequence of Events

heterogeneity in bank-specific (or market-specific) opportunities. The distribution of manager types will describe the investment opportunities available in the banking industry. Manager type is private information. For simplicity all banks are assumed to have the same leverage and cost of funds.[11]

The timing of the model is shown in Figure 15.2. At date 1 bank managers choose a loan portfolio. Outside owners (outsiders) observe the type of portfolio, but not its quality (the manager's type). At date 2 outsiders observe the outcomes of managers' loan choices. At this time outsiders may decide to fire some managers, but this is costly. If a manager is fired, shareholders have two choices at date 2. They can replace the fired manager with a new bank manager and continue investing in the banking industry. Or, they can move resources into a nonbanking investment.[12] Finally, also at date 2, new loans or other investments are made which have realizations at date 3. Managers receive private benefits, *w,* in each period that they are in control of the bank. If managers are fired by the outside owners at date 2, they earn no control rents at date 3. All agents are risk neutral.[13]

We look for a Sequential Nash equilibrium: a firing decision rule that maximizes the utility of outsiders given the lending decisions of each type of manager

11. The effects of deposit insurance will be discussed in a subsequent section.

12. We assume that the cost of funds and leverage are again the same for all banks at date 2.

13. Risk neutrality is the simplest assumption and possibly the most realistic. The realism of the assumption depends on the extent to which individual managers have plunged in bank

and a date 1 lending decision rule for each type of manager that maximizes utility given the outsiders' rule for firing managers.[14]

In specifying the loan portfolio available to managers, we want to parsimoniously contrast the decisions outsiders want managers to make and the decisions managers do, in fact, make. Thus, we need to include portfolio choices where managers might prefer a riskier choice than outsiders and vice versa. For this, we need four types of loan portfolios—"good" and "bad" risky portfolios as well as "good" and "bad" safe portfolios. Assume that a risky loan portfolio has a bivariate return, either R or 0. What differentiates a good risky portfolio from a bad one is the probability of getting a high return. A good risky portfolio returns R with probability θ_G while a bad risky portfolio returns R with probability θ_B, where $\theta_G > \theta_B$. Assume that a safe loan portfolio yields its expected value with probability one. A good safe loan portfolio has a return S_G and a bad safe loan portfolio has return S_B, where $S_G > S_B$. Safe loan portfolios are meant to include such assets as consumer loans and home mortgages. Assets such as Treasury bills and bonds, while possibly a part of a safe loan portfolio, offer similar yields to all types of bank managers, and thus do not serve to separate managerial types in the manner we want.

There are two cases that exemplify why managers' and outsiders' preferences might differ. The first case is when managers choose between a good risky portfolio and a good safe portfolio, where the risky portfolio offers a higher expected return than the safe portfolio:

Assumption 1. *The expected value of a good risky loan portfolio is greater than the expected value of a good safe loan portfolio:* $\theta_G R > S_G$.

We refer to managers that choose between good safe and good risky portfolios as "good" managers. This is the traditional problem examined in the corporate control literature in the sense that the industry is not declining. The second case is when managers choose between a bad risky portfolio or a bad safe portfolio, where the risky portfolio offers a lower expected return than the safe portfolio:

Assumption 2. *The expected value of a bad risky loan portfolio is less than the expected value of a bad safe loan portfolio,* $S_B > \theta_B R$.

We refer to managers that choose between bad safe and risky portfolios as "bad" managers. One interpretation of these manager types is that good

stock when we allow for them to own bank stock later in the paper. Williams (1987) considers the interaction between risk aversion and incentives when there are agency problems in firms.

14. Sequential Nash equilibrium also requires that beliefs satisfy a consistency requirement. As will be seen, this is straightforward in our model.

managers are those that can adapt to technological changes while bad managers cannot adapt.

It is important to emphasize that underlying our model is the existence of other types of managers that always choose the first-best portfolio. That is, for any firing rule that outsiders use, the other types of managers make the portfolio choices, either risky or safe, that their outsiders want them to make. Two types in particular are necessary. Assume that some managers are only able to invest in risky portfolios (or that the safe portfolios available to them offer a significantly lower expected return than the risky portfolios). Some of these managers invest in good risky portfolios and others invest in bad risky portfolios.[15]

The dividing line between a good loan portfolio and a bad loan portfolio is the point at which an outsider is indifferent about whether to fire managers if they knew the quality of the loan portfolio. In deciding whether or not to fire a manager, outsiders compare the expected return on their investment in the bank to the alternatives of hiring another bank manager or investing in a nonbanking alternative. The outsiders must also incur a cost, c, to fire the current manager (more generally, there is a liquidation cost for capital which includes firing costs; this cost is assumed to be borne by the bank). Clearly, a manager is fired if the expected increase in return from either hiring a new manager or investing in a nonbank alternative exceeds the cost of firing the manager. Let Γ be the return from the nonbanking alternative and let V be the expected return from banking if a new manager is hired (net of the private benefits, w). Then, the opportunity cost of retaining a particular manager is:

$$X = \text{Max}[V, \Gamma] - c.$$

The parameter Γ is exogenous as is V (since V depends on the relative proportions of different manager types). Note that $V < \Gamma$ would mean that there is overcapacity in the banking system, that is, the (expected) return on the nonbanking alternative, Γ, is higher than the expected return in banking. This occurs when the number of bad managers is relatively high. As a result, bank equityholders would prefer to move their resources out of banking at date 2 when they fire a manager. Below we discuss the relationship between V and Γ further.

Assume that the expected return is such that outsiders, conditional on knowing a manager's type, fire managers that have only bad investment opportunities and not managers that have at least one good investment opportunity. This assumption is stated as:

15. Note that the focus on good managers, defined by Assumption 1, and bad managers, defined by Assumption 2, does not preclude the presence of managers with opportunities such that $\theta_G R < S_G$ or $S_B < \theta_B R$. It is easy to introduce a number of other types of managers. Adding other types does not change any of the results (see Gorton and Rosen (1992)).

Assumption 3. *Outsiders want to fire only bad managers (those that have a choice between a bad risky loan portfolio and a bad safe loan portfolio):* $S_G - w > X > S_B - w$.

This condition is sufficient for any set of portfolio opportunities, by Assumptions 1 and 2.

Below we investigate the optimality of various rules for firing managers that could be adopted by outside shareholders. Throughout, however, we will assume that the costs of firing a manager are small enough that outsiders fire any manager that chooses a bad safe loan portfolio, because that manager is revealed to be a bad manager. This assumption is not crucial. It is important that outsiders are unable to determine the type of manager that chooses a risky project from ex post returns (since successful risky projects earn R, but the ex ante probability of earning R is not observed).

15.1.2. Preliminary Analysis

To see how private benefits affect managerial choices, suppose for illustrative purposes that the outsiders fire bad managers that choose safe loan portfolios (their quality is revealed by the realization) along with managers that choose risky loans and earn zero. By assumption, outside shareholders want good managers to choose risky loans (Assumption 1) and bad managers to choose safe loans (Assumption 2). Of course, managers take their private benefits into account when they evaluate loans. If good managers make risky loans, then there is some chance that they are fired. On the other hand, if good managers make safe loans they are never fired. Thus, because of the private benefits, good managers choose safe loans and behave too conservatively (when we say a portfolio choice is "too conservative" or "too risky" we always mean relative to first-best). Bad managers are in the opposite situation from good managers. If they choose safe loans, they are fired, but if they choose risky loans and get a high return, they retain their job. This leads bad managers to choose risky loan portfolios.[16]

By explicitly modeling both good and bad managers, we are able to characterize the state of the industry. This is important because the aggregate behavior of the industry depends on the relative proportions of different manager types. In the existing literature, the implicit assumption is that good managers predominate. In that case, the conservatism of good managers drives the aggregate level

16. For this to be an equilibrium, the assumed firing rule of the outsiders must be a best response to the lending strategies. This depends on the relative proportion of good managers to bad managers and on the firing cost. We omit this calculation here.

of risk-taking. On the other hand, if, as we assume, there is a high proportion of bad managers, then aggregate investments reflect the risky decisions of the bad managers.

Managerial entrenchment occurs when outsiders are unable to determine whether their manager is taking a first-best action or when it is too costly to fire a manager. In the example above, managers make suboptimal choices because outsiders are unable to distinguish manager type based on the return to risky portfolios. Implicit in the analysis above is the assumption that the firing cost, c, is low enough that outsiders want to fire managers that choose risky portfolios and get a return of zero. If the firing cost is large enough, the outsiders may find it optimal to retain managers that earn zero on risky loans. This would be a more extreme form of entrenchment.

15.1.3. Managerial Ownership

When managers are shareholders in the firms they manage, the situation is more complicated than the preliminary analysis above because managers not only receive private benefits from managing, but also benefit from ownership of a (publicly observable) fraction, α, of the stock in the bank. Ownership influences portfolio choice because decisions taken to maintain private benefits can reduce the value of the stock.

Managerial ownership of banks can affect the outsiders' cost of firing managers. The decision to fire the manager is made by the board of directors. Board membership control (by managers) is likely to depend on managerial stock ownership. Also, to the extent that managers own stock they can demand such things as larger severance pay, making firing more costly. We assume that the cost of firing a manager is increasing in the manager's ownership share, $c(\alpha)$. If firing is too expensive, then owners would prefer to bear the cost of a bad manager rather than pay the firing cost. A sufficient bound on the firing cost which ensures that bad managers are not retained solely because the cost of firing is prohibitive is given by:

Assumption 4. $c'(\alpha) < w/\alpha^2$.

(This assumption reappears in the proofs in Appendix 1.) We also assume that, if fired, managers still receive the value of their shares at date 3. Note that since the final date is the end of the model, if a manager is not fired, the date 2 portfolio choice is straightforward: the manager, being a shareholder, simply chooses the first-best portfolio.

In the preliminary analysis discussed briefly above, risk-taking in the banking industry depends only on the relative proportions of good and bad managers

and the firing cost. When managers own stock, however, overall risk-taking in banking also involves the distribution of stock ownership across manager types.

Rather than go through the model in detail, we provide an overview of the results. (Details of the model, and proofs of the propositions, are presented in Appendix 1). Recall that the costs of firing a manager are assumed to be small enough that outsiders fire any manager that chooses a bad safe loan portfolio because that manager is revealed to be a bad manager. However, outsiders are unable to determine the type of manager that chooses a risky project from ex post returns. Thus, any firing rule they use inevitably allows either bad managers to continue or good managers to be fired. There are three firing rules outsiders could adopt toward managers that choose a risky loan portfolio: (a) fire all managers that earn a low return of zero on their risky loan portfolio; (b) fire no managers that choose a risky portfolio; (c) fire all managers that choose a risky loan portfolio. Finding the equilibria of the model is essentially a process of examining the responses of managers to each firing rule. Since managerial ownership is observable, the firing rule depends on managerial ownership.

In what follows, we concentrate on the conditions under which (a) is the equilibrium firing rule for all levels of managerial ownership. Throughout the discussion, bear in mind that if firing costs are high enough, firing rule (b), not firing rule (a), will be the equilibrium. Clearly, when firing rule (b) is selected by outsiders, bad managers are entrenched because their jobs are protected when they choose the risky, second-best, portfolio. It is straightforward to show that for a given managerial ownership share, options (b) and (c) can only be equilibria if the proportion of managers that can choose a bad risky loan portfolio (whether or not it is the first-best) is, respectively, low enough or high enough relative to the proportion of managers that can choose a good risky loan portfolio. Sufficient conditions for (a) to be optimal are given below.

The equilibrium choice of a lending strategy by good and bad managers involves the trade-off among three factors: the private benefits of working at date 2, the cost to the manager as a shareholder from any non-expected-value maximizing choice of a loan portfolio at date 1, and the cost of firing the manager. At low levels of managerial ownership, private benefits are more important to managers than their ownership share. For higher levels of managerial ownership, managers place more weight on bank return and less on private benefits. In the limit, when the manager owns the entire bank, only the bank return matters. So:

PROPOSITION 1. *Assume Assumptions 1, 2, and 3 hold and outsider owners fire all managers that earn a low return of zero on their risky loan portfolio (firing rule (a)). Then there exists an ownership share* α^* *such that good managers choose safe loans if and only if* $\alpha \leq \alpha^*$. *There exists an* α^{**} *such that bad managers choose risky loans if and only if* $\alpha \leq \alpha^{**}$.

The proposition says that good managers, who choose risky loans in the absence of agency costs, choose safe loans if their equity stake is lower than a critical level, α^*. Bad managers, who choose safe loans in the absence of agency costs, instead choose risky loans if their equity stake is lower than a critical level, α^{**}. In other words, if managerial equityholding is not high enough to align managers' incentives with those of outside equityholders, then managers deviate from first-best portfolio choice. The proposition identifies the level of managerial shareholding at which this change occurs. Moreover, the deviation depends on whether the manager has good or bad investment opportunities and on the firing cost.

The optimality of firing rule (a) depends on the cost of firing a manager and the proportions of manager types at any given level of managerial ownership. We can find a set of sufficient conditions to ensure that firing rule (a) is used:

PROPOSITION 2. *Assume Assumptions 1, 2, and 3 hold. Then there exists a unique equilibrium for any managerial ownership level, α, in which outsiders choose to fire all managers that earn a low return of zero on their risky loan portfolio (firing rule (a)), and managers behave as described in Proposition 1, if the following two conditions hold:*

$$\frac{\gamma_B}{\gamma_{GG}+\gamma_G}\,\frac{(X+w-S_B)}{(\theta_G R-X-w)} \geq \frac{1-\theta_G}{1-\theta_B} \tag{15.1}$$

$$\frac{\theta_G}{\theta_B} \geq \frac{\gamma_{BB}+\gamma_B}{\gamma_G}\,\frac{(X+w-\theta_B R)}{(\theta_G R-X-w)} \tag{15.2}$$

where γ_{GG} is the proportion of good managers; γ_{BB} is the proportion of bad managers; γ_G is the proportion of managers that always choose a good risky loan portfolio; and γ_B is the proportion of managers that always choose a bad risky loan portfolio $(\gamma_{GG}+\gamma_{BB}+\gamma_G+\gamma_B=1)$.

The two conditions in Proposition 2 characterize when it is optimal to fire all managers that earn a return of zero on their risky loan portfolio. The conditions are not restrictive, that is, it is not the case that the proportion of bad managers need be very large for this equilibrium to exist. For example, suppose $R=1, \theta_G = 0.9, \theta_B = 0.6, S_G = 0.8, S_B = 0.7$, and $X+w=0.75$. Then the conditions of the proposition require that $\gamma_B/(\gamma_{GG}+\gamma_G) \geq 1/12$ and $(\gamma_{BB}+\gamma_B)/\gamma_G \leq 3/2$. These conditions are satisfied, for example, by: $\gamma_G = 0.3, \gamma_{GG} = 0.3, \gamma_B = \gamma_{BB} = 0.2$. Another example satisfying the conditions is: $\gamma_G = \gamma_{GG} = 0.4$ and $\gamma_B = \gamma_{BB} = 0.1$.

The two conditions of Proposition 2 also can be used to illustrate the conditions under which the other firing rules would be optimal. In particular, if condition (15.1) does not hold when S_B is replaced by $\theta_B R$ and condition (15.2) holds (roughly, too few good managers), then outsiders want to fire any

managers that choose a risky loan portfolio. Conversely, if condition (15.1) holds and condition (15.2) does not hold when $\theta_B R$ is replaced by S_B (too many good managers), then outsiders do not fire managers choosing risky portfolios.

The equilibrium conditions in Proposition 2 depend on the cost of firing, $c(\alpha)$, since the firing cost is embedded in the opportunity cost of firing a manager, X. As the firing cost increases, outsiders find it less profitable to fire a manager, even if the manager makes risky loans and earns a zero return.

15.1.4. Equilibrium Managerial Entrenchment

An important feature of the equilibrium described by Proposition 2 is that not all bad managers are detected and fired at date 1. Bad managers that choose risky loan portfolios and have a high payoff (of R) continue to make loans at date 2. This is because these bad managers have successfully pooled with the good managers. The frictions caused by asymmetric information and costly firing prolong the period during which these managers are left in control of their banks. This persistence can explain why the banking industry appears to have adjusted slowly to the changed investment opportunities, since changed opportunities are captured here by the relatively high proportion of bad types.

Our goal is to find the aggregate pattern of risk-taking in the industry as a function of the equity ownership structure of banks (in cross-section). This relationship is likely to be highly nonlinear because it depends on the distribution of manager types and on the distribution of insider holdings across these types. Proposition 2 provides sufficient conditions for existence and uniqueness of an equilibrium with managerial entrenchment. But, to be more precise, we need to know the relationship between the critical ownership shares at which good and bad managers switch from second-best to first-best portfolio choices (α^* and α^{**} in Proposition 1). The critical levels α^* and α^{**} are determined by the tradeoff between the lost private benefits in period 2 when the manager is fired for taking the first-best action and the gain in the return on the manager's stock from taking the first-best action. Good managers that choose risky portfolios are fired only when they are not successful (and earn zero). If it is very probable that a risky portfolio is successful, then a good manager has little to fear from choosing the first best. We can show:

PROPOSITION 3. *Assume Assumptions 1–4 hold and outside owners fire all managers who earn a return of zero on their risky loan portfolios (firing rule (a)). Then:*

$$\theta_B (\theta_G R - S_G) + (1 - \theta_G) ((1 - \theta_G) \theta_B R - (1 - \theta_B) S_B) > 0, \qquad (15.3)$$

implies $\alpha^* < \alpha^{**}$. *Further, if conditions (15.1) and (15.2) of Proposition 2 hold, then there is a unique equilibrium with* $\alpha^* < \alpha^{**}$.

Condition (15.3) of the proposition holds when the expected return on good risky loans is "high" (as $\theta_G \to 1$, (15.3) holds for any values of the other parameters). Since this is unobservable we cannot test it directly. Nevertheless, Proposition 3 provides an illustrative characterization of the pattern of aggregate risk-taking in an unhealthy banking industry that we use as a null hypothesis in our empirical work.

Note that condition (15.3) holds for the examples given after Proposition 2. Figure 15.3 illustrates the pattern of aggregate risk-taking for the first example. It shows that, over the range of managerial ownership between 0 and α^*, bad managers choose risky portfolios and good managers choose safe portfolios (and all other types of managers choose their first-best portfolios). Between α^* and α^{**}, both good and bad managers choose risky portfolios (and, again, all others choose their first-best portfolios). Above α^{**}, bad managers choose safe portfolios and good managers choose risky portfolios (and all others choose the first-best). Figure 15.3 provides a concrete example showing how entrenched managers can distort aggregate risk-taking.

Figure 15.3, drawn under the assumption that banking is dominated by a lack of good lending opportunities, also illustrates a major difference between our

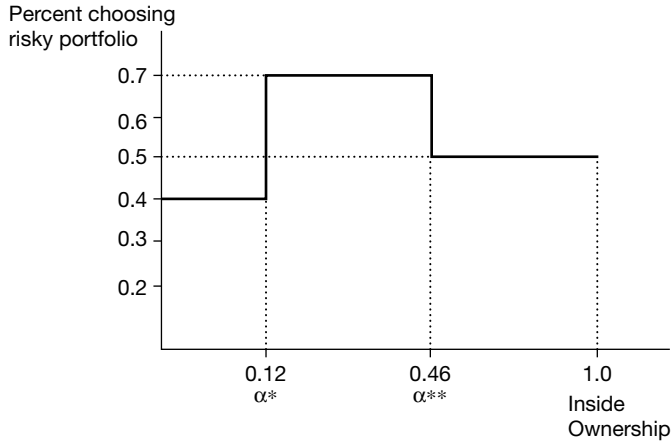


Figure 15.3 Example of Aggregate Risk-Taking. The example assumes $R = 1.0, \theta_G = 0.9$, $\theta_B = 0.6$, $S_G = 0.8$, $S_B = 0.7, \gamma_G = 0.3, \gamma_{GG} = 0.3$, $\gamma_B = 0.2$, and $\gamma_{BB} = 0.2$. Using these values α^* and α^{**} can be calculated as can the optimal decisions of each manager type. For values of α between zero and α^*, bad managers choose risky portfolios and good managers choose safe portfolios (and all other manager typos choose their first-best portfolios). Between α^* and α^{**}, both good and bad managers choose risky portfolios (and others choose first-best, as before). For values above α^{**}, bad managers choose safe portfolios and good managers choose risky portfolios

model and other models of corporate control. Like other corporate control models, we find conditions under which managers take second-best actions. But, as the figure illustrates, when there are sufficient bad managers in an industry, the traditional result that corporate control problems lead to excess conservatism on the part of managers is reversed.

15.2. DISCUSSION OF THE MODEL

In this section we briefly discuss the main assumptions of the above model. The assumptions discussed are as follows. The model does not have debtholders or regulators playing an active role. Also, it uses a simple ownership structure for both insiders and outsiders. Finally, we have identified investment opportunities in banking with manager types rather than with inherent characteristics of particular banks, independent of the manager.

15.2.1. Debtholders and Bank Regulators

The analysis assumes that bank depositors continue to deposit one dollar in each bank in the banking industry despite the fact that there are many bad managers. We justify this assumption for banks by appealing to (fixed-rate) deposit insurance. Deposit insurance allows banks to raise funds even when many bank managers are bad. Since the interest paid to depositors is independent of managers' actions, there is no reason for insured depositors to become informed. Further, insured and uninsured depositors face the same information problems that outside shareholders do. Allowing debtholders to play an active role (without deposit insurance) would reduce the return to the risky activity because debtholders would demand higher interest rates. But, the qualitative results of the model would not change.

The model assumes outside shareholders have no opportunity to produce information about manager types at date 1. Such information could allow outsiders to make more refined firing decisions. We consider this possibility in Gorton and Rosen (1992). When monitoring, i.e., producing information about manager type at date 1, is possible but costly, the essential features of the equilibrium remain unchanged. In particular, if outsiders monitor managers that choose risky loan portfolios and earn zero (and do not monitor managers that earn R on risky portfolios), then the only difference from the basic model is that good managers need not fear earning zero on risky portfolios. But, the incentives of bad managers are unchanged; they are fired unless they choose risky loan portfolios and earn R.

The model also assumes that outside shareholders act as a single agent. Since outside shares are often widely dispersed, possibly causing a free rider

problem in monitoring and firing, the presence of a few block shareholders may be important for initiating monitoring and firing.[17] Firing and monitoring costs may depend on the fraction of outside shares that are held in blocks. Blockholders should reduce firing and monitoring costs. We include this consideration in the empirical work below. It has straightforward implications for the above analysis.

We have also not considered the role of bank regulators. Regulators might examine banks (monitor) and close banks (fire managers) under different circumstances than outside shareholders do. As discussed in Gorton and Rosen (1992), if outside shareholders face very high monitoring costs, then they do not monitor, but instead fire managers based only on loan returns. Regulators may face lower monitoring costs than outsiders, leading to most monitoring being done by regulators.

Government regulators, in addition, have more power than private citizens. In particular, they can examine banks ex ante and impose ex ante restrictions on risk-taking. Also, regulators can impose punishments ex post, such as banning individual bank managers from working in the banking industry. To the extent that they are costless, and that regulators face the right incentives, these actions can mitigate the problems we analyze. Others, however, argue that agency problems between regulators, Congress and the public distort regulators' incentives. (See, e.g., Kane (1992).)

15.2.2. The Equity Ownership Structure

Like previous researchers in this area, we assume that the distribution of equity ownership is given and, in particular, that bank managers own bank equity. This is important in our model because equity shares have voting rights and we have related this to firing costs (by assuming that these costs to outsiders are increasing in the fraction of shares owned by management). We provide no reason why managerial compensation should be in the form of equity shares with voting rights.[18] Obviously, in a larger model the equity ownership structure would have to be endogenized and this is a subject of further research. For our purposes managerial stockholdings are given.

A related issue concerns compensation in general. Managers that at date 1 know, privately, that they are good might accept a different compensation package than bad managers. That is, a separating equilibrium might exist. The agency

17. See Shleifer and Vishny (1986). The empirical evidence supports the importance of large shareholders in increasing firm value. See Mikkelson and Ruback (1985), Holderness and Sheehan (1985), Barclay and Holderness (1990), and Zeckhauser and Pound (1990).

18. Gorton and Grundy (1995) provide an argument for why firms would find it optimal to reward managers with voting equity.

problem we focus on can be mitigated to the extent that compensation contracts for managers can be designed to align their interests with those of outside shareholders. Of course, it may be that managers learn about the decline in investment opportunities after such contracts have been signed. In addition, as discussed below, the interpretation of types as corresponding to managers, rather than to banks, is only a simplification. Compensation contracts in banking is another area for further research.[19]

15.2.3. Investment Opportunities and Overcapacity in Banking

Intuitively, the conditions in Proposition 2 say that, ceteris paribus, the equilibrium depends on the return to an investment made by the current manager, given the relative proportions of good and bad managers, compared to the alternative, X (recall that $X = \text{Max}[V, \Gamma] - c(\alpha)$). While the model takes Γ as exogenous, its role is important. If the expected value of the bank, conditional on drawing new managers from the population of managers at date 2, V, is less than the value of investing in the nonbanking alternative, Γ, then resources will leave the banking industry at date 2. The banking industry is unhealthy when bad managers are relatively common, causing the expected value of an investment in banking (by an outsider) to be low (relative to the alternative). If the banking industry is so unhealthy that outside shareholders would prefer to invest their resources in the nonbanking alternative at date 2, then there is overcapacity in the banking industry ($V < \Gamma$).

While it might be natural to assume that the conditions of Proposition 2 correspond to overcapacity in the banking industry, the model does not, strictly speaking, allow us to make that statement. However, that is an artifact of how investment opportunities are modeled. We modeled investment opportunities as corresponding to the distribution of manager types with different lending choices. An alternative interpretation is consistent with the results. Instead of managers being of different types, we might imagine that the banks themselves face different investment opportunities and that all managers are the same. In this case there is no alternative of hiring a different manager to obtain better

19. Compensation contracts in banking have been studied by Boyd and Graham (1991), Mullins (1993), Houston and James (1993), and Booth (1993). Boyd and Graham (1991) find that in banking, management compensation is positively, and significantly, related to asset size, but not significantly related to profitability. Mullins (1993) finds that bank managers' salaries and stock options are not related to risk-taking (as measured by the standard deviation of stock returns). Houston and James (1993) find no evidence that bank compensation is structured to induce risk-taking, but is related to measures of growth opportunities. Booth (1993) finds that the determinants of bank CEO compensation are similar to those of nonfinancial firms, except that bank managers' total compensation is more sensitive to board members' stock ownership.

performance, so poor investment opportunities means that $V < \Gamma$. Consequently, outside shareholders will want to fire the managers of bad banks since they prefer to move their resources out of banking. Managers of bad banks will want to avoid this because they will be out of jobs. Since the industry is shrinking (i.e., $V < \Gamma$), they will not be rehired at another bank. Thus, this interpretation is consistent with the above results and implies that there is overcapacity in banking.

15.3. EMPIRICAL IMPLEMENTATION OF THE MODEL

Our goal is to test the corporate control model against the alternative hypothesis of moral hazard. Towards that end, in this section we first explain how the two views can be distinguished. Then, in order to conduct the tests, we empirically determine which categories of loans correspond to the predictions of the model in terms of risk and return characteristics. (Test results are reported in Section 15.4.)

15.3.1. Hypotheses

Proposition 3 allows us to test the joint hypothesis that corporate control problems are important in bank portfolio choice and that the industry is unhealthy. We can look for a pattern of risk-taking in the data that is similar to Figure 15.3. The proposition implies that the pattern of risk-taking as a function of managerial ownership is inversely U-shaped, rising and then falling. But, the nonlinearity may be more complicated since the model has discrete manager types and discrete choices. Nevertheless, and this is the main point, the model allows us to distinguish our hypothesis from the leading alternative hypothesis of moral hazard due to fixed-price deposit insurance. In particular we can test:

HYPOTHESIS 1. *Over some intermediate range of insider ownership, the relationship between risk-taking and the share of insider stock ownership, α, is inversely U-shaped.*

Notice that if there were a sufficient proportion of good types in the banking industry, we would predict a U-shaped relationship between risk-taking and managerial ownership.

The leading alternative hypothesis to the corporate control arguments outlined above is the moral hazard hypothesis. Moral hazard models concentrate on the conflict between banks and regulators. Bank managers' interests are assumed to be aligned with those of the bank owners. In the canonical moral hazard model, the banking industry is unhealthy in the sense that charter values have

declined (e.g., Keeley (1990)). Owners attempt to take advantage of fixed-rate deposit insurance by making relatively risky portfolio choices. In this theory, there is no predicted relation between risk-taking and the fraction of bank stock held by bank managers, α. Thus, one alternative hypothesis is:

HYPOTHESIS 2. *There is no relationship between managerial ownership, α, and risk-taking.*

More charitably, one might suppose that the moral hazard model applies when outside shareholders can control bank managers. This could occur if the manager's fraction of stock is low or very high. Low levels of insider holdings increase the ability of outsiders to control managerial decisions, and high levels of insider holdings mean that managers' interests align with those of outsiders. So, moral hazard models might be interpreted to predict that owner-controlled banks, and perhaps banks with low levels of insider ownership, make relatively risky portfolio choices compared to banks with entrenched managements:

HYPOTHESIS 3. *Above some level of managerial ownership, risk-taking is increasing in α. At low levels of insider ownership, risk-taking may be decreasing in α.*

Corporate control and moral hazard predict sharply different patterns of risk-taking in an unhealthy banking industry. Our corporate control model predicts that risk-taking is inversely U-shaped with respect to managerial ownership. Moral hazard models predict either no relation or the opposite: either risk-taking is U-shaped with respect to α (or it is increasing above a certain point).

In a more general model, fixed-rate deposit insurance, through its negative effect on monitoring by bank depositors, also can influence bank risk in ways that are independent of insider ownership. The absence of active monitoring of banks by depositors may reduce the incentives of bank managers to put in effort to screen potential borrowers. Thus, to the extent that bank shareholders do not want their managers spending extra time screening borrowers, fixed-rate insurance increases the overall risk in banking. This is a type of moral hazard. But, more commonly, bank owners and bank depositors have a similar interest in encouraging monitoring of borrowers by managers. When interests coincide, the pattern of risk-taking by managers should be a function of corporate control problems, not moral hazard.

15.3.2. Risk, Return, and the Composition of Banks' Loan Portfolios

As a first step toward testing our predictions on portfolio choice by bank managers, we divide bank loan portfolios into categories that are relatively risky and

relatively safe. In the next section, we investigate how portfolio composition is related to the pattern of equity ownership.

What we would like is to provide evidence of the ex ante risk and return characteristics of bank loan portfolios. Unfortunately, it is not possible to determine what bank managers think the expected return on a loan portfolio is. Instead, we are forced to use ex post data from bank *Call Reports of Income and Condition* for year-end 1984–1990. The risk of a bank portfolio is estimated by using the proportion of loans that are nonperforming. (Nonperforming loans are those that are 90 days or more past due or not accruing interest.)[20] By this measure, the risk of bank loans rose considerably in the 1980s. Panel A of Table 15.1 shows a breakdown of nonperforming loans by loan category. Commercial and Industrial loans (C&I loans) are the riskiest and consumer loans are the safest. The average real estate loan lies somewhere in the middle, but this category includes different types of loans.

Since the risk figures for real estate loans aggregate loan categories that we would expect to be (relatively) safe (such as home mortgages) with categories that are possibly very risky (such as construction and development loans), we need to find a way to disaggregate real estate loan risk. We have 1991 and 1992 data on nonperforming real estate loans by loan type. For banks over $300 million in assets, 7.9 percent of real estate loans were nonperforming. Construction and development loans had a nonperforming rate of 20.3 percent; commercial loans had a nonperforming rate of 10.1 percent, and mortgages had a nonperforming rate of 3.1 percent. Thus, construction loans and commercial loans were both riskier than C&I loans and consumer loans. We expect that the pattern in 1991 and 1992 is representative of the pattern in the 1984–1990 period, although we recognize that 1991 and 1992 were bad years for construction and commercial real estate loans.

Examining the return on bank loans provides evidence that banking was unprofitable in the 1980s. Panel B of Table 15.1 gives the return on loans (ROL) for banks over $300 million in assets. The first column is the gross ROL, while the second column presents the ROL net of the average interest rate on deposits. The average interest rate is deducted from the ROL in an attempt to measure the net return on bank loan portfolios. As the table shows, the gross ROL (column 1) has fallen, but some of the decline occurred at the same time as a decline in interest rates. The ROL net of the average interest rate (column 2) also fell, but by less than the gross ROL.

20. The risk of a loan should be evaluated by the contribution of the loan to overall bank risk, but data limitations prevent this computation. Thus the risk of each category of loans is evaluated independently. The implicit assumption is that no category of loans contributes significantly more than any other to the diversification of bank's return stream. We also ignore interest rate risk due to data limitations.

Table 15-1. RISK AND RETURN ON BANK LOANS, 1984–1990 (BANKS OVER $300 MILLION IN TOTAL ASSETS)

Panel A shows the fraction of loans that are nonperforming, by loan type. Nonperforming loans are loans that are more than 90 days past due, nonaccruing loans, and other real estate owned (foreclosed real estate). Panel B shows the return on loans (interest income on loans divided by total loans) and the return net of the average interest rate paid on deposits (net interest expense divided by total deposits and other interest-paying liabilities). Panel C shows the difference between the return on various loan categories and the average return on all loans. The source for all data is the *Call Reports of Income and Condition.*

Panel A: Rate of Nonperforming Loans, by Loan Type

Year	Total Loans	All Real Estate	C&I Loans	Consumer Loans
1984	2.71	2.81	5.38	1.53
1985	2.64	2.72	4.79	2.17
1986	2.97	3.27	4.96	2.62
1987	4.63	3.60	6.86	2.82
1988	4.15	3.09	5.33	2.71
1989	4.48	4.05	5.30	2.92
1990	5.66	6.38	6.94	3.47

Panel B: Return on Bank Loans

Year	Return	Return Net of Average Interest Paid
1984	11.23	2.01
1985	10.19	2.35
1986	8.74	2.20
1987	8.74	2.11
1988	9.28	2.01
1989	10.29	1.62
1990	9.67	1.41

Panel C: Additional Return on Bank Loans Above Average for All Loans, by Loan Type

Year	Net Additional Return on All Real Estate	Net Additional Return on All C&I Loans	Net Additional Return on All Consumer Loans
1984	–0.83	0.02	1.09
1985	–0.40	–0.48	2.12
1986	–0.30	–1.09	2.89
1987	–0.70	–0.51	2.47
1988	–1.02	–0.64	1.74
1989	–1.28	–0.24	1.27
1990	–1.23	–0.54	1.88

For a risky loan to be a bad gamble for an entrenched manager, the loan must offer a lower expected return than safer loans. A direct estimate of the return on the categories of bank loans is possible for C&I loans, consumer loans, and (total) real estate loans. To show the relative return for the different loan categories clearly, Panel C of Table 15.1 presents the difference between the return on each loan and the average return on all loans. The return on C&I loans and on real estate loans are below average, while consumer loans get an above average return.

Of course, one explanation of the risk and return characteristics discussed above is bad luck. If bad luck caused the low return and high risk of real estate construction and development loans, then there should be no relationship between this type of lending and managerial ownership. Our results suggest that if corporate control problems are important, bad entrenched managers should make the most real estate construction loans and the fewest consumer loans, with C&I loans somewhere in between. We concentrate on these three loan categories.

15.4. INSIDERS AND OUTSIDERS IN BANKING: TESTS

In this section we test the hypothesis that when the banking industry is unhealthy, banks with entrenched management invest in the relatively risky commercial real estate construction and development loans and less so in the relatively safe category of consumer loans.

15.4.1. Data on Equity Ownership

In order to distinguish between moral hazard problems and corporate control problems, we collect data on the ownership structure of bank holding companies. Ownership data are a cross-section of holdings in 1987/88 as described in Appendix 2. We use two measures of ownership, the holdings of insiders (directors and officers of the bank) and the holdings of outsiders (that is, noninsiders) that hold at least five percent of the outstanding stock.[21] Our measure of outside concentration includes large blockholders and serves as a proxy for the degree of outsider control. Panel A of Table 15.2 provides summary measures of our data together with the summary measures for nonfinancial firms provided by McConnell and Servaes (1990). Outsider concentration in nonfinancial firms is larger than in banks. The same is true for insider holdings.

21. Data from SEC 10-K reports require that shareholders with at least five percent holdings report their holdings, but the holdings of others with less than five percent are also sometimes reported.

Table 15-2. INSIDE AND OUTSIDE SHAREHOLDERS OF BANKS AND NONFINANCIAL FIRMS

The data on bank holding companies in Panels A and B come from SEC filings (see Appendix 2). The data on nonfinancial firms in Panel A are from McConnell and Servaes (1990). Insiders are Board members and family of Board members. Outsiders are other shareholders with at least five percent ownership.

Panel A: Summary Statistics on Insider and Outsider Holdings

	Bank Holding Companies	**Nonfinancial Firms**
Sample size	458	1,093
Average Insider Holdings (%)	15.25	11.84
Median Insider Holdings (%)	8.33	5.00
Range of Insider Holdings (%)	0–99	0–89
Average Outsider Holdings (%)	7.87	25.60

Panel B: Proportion of Banks in Sample, by Share of Insider Ownership

Share (%)	**Number of Banks**	**Proportion of Banks (%)**
Less than 5	166	36
5–10	84	18
10–25	107	24
25–50	71	16
Greater than 50	30	7
Total	458	100

15.4.2. The Estimation Procedure

Our goal is to empirically analyze the relationship between the share of particular loan types (of total assets) and the share of the firm held by insiders. In order to estimate and draw inferences some structure must be imposed on the relationship. This issue of functional form seems particularly important since Morck, Shleifer, and Vishny (1988) and McConnell and Servaes (1990), studying nonfinancial firms, obtain essentially contradictory results using two different ad hoc nonlinear parametric specifications, while, for banks, Saunders, Strock, and Travlos (1990) use a linear specification.

Looking at Panel B of Table 15.2 conveys some sense of the difficulties. Panel B of Table 15.2 shows that over one-third of the banks in our sample have insider ownership of less than five percent. Nonfinancial firm samples also have a large number of observations at less than five percent insider ownership. Above five percent observations on insider holdings are more sparse. This suggests that the results of estimating almost any parametric specification would almost certainly

be driven by managers with very small ownership shares.[22] It is quite likely that many parametric specifications would result in "significant" coefficients, though they might well not be consistent estimates.

Thus, although our model predicts, under the conditions of Proposition 3, that over some range of managerial ownership, the relationship between risky lending and managerial ownership is inversely U-shaped, estimating a quadratic relationship over the entire range of ownership shares could provide misleading results.

For these reasons, our empirical analysis is in two parts. We begin by imposing as little structure as possible, and then move on to imposing more structure. The first approach imposes no a priori functional form on the relationship between insider ownership and portfolio choice. In particular, this procedure does not impose a quadratic specification a priori. Nonparametric methods can uncover the exact nonlinear relationship (at least asymptotically) between the particular loan share choice and insider holdings. Of course, using a nonparametric procedure to estimate the relationship between insider holdings and portfolio choice, we also want to control for a number of other factors which can be expected to affect the relationship. This motivates our semiparametric procedure.

The semiparametric procedure has less precision than parametric models. The trade-off between the larger standard errors of the semiparametric model and the possibly incorrect inferences of the parametric model, discussed further below, leads us to impose further structure based on the first set of results. In particular, we also use a quadratic specification to check for the inverse U-shape predicted by Proposition 3, but with the quadratic specification we restrict attention to an intermediate range of insider holdings.

Let $\mathbf{L_i}$ be the vector with elements consisting of the fraction of loan type i in the total bank portfolio of a sample of banks.[23] Let α be the vector of insider fractional holdings. Also define the following variables: the vector **O** has elements consisting of the fraction held by outside block shareholders in each bank; the vector of the log of total assets in each bank is **A**; the loan to total assets ratio is **N**; **Yr** indicates dummy variables for the year; **Z** indicates the region of the country in which the bank operates.[24] Letting the matrix **X** be the matrix consisting of

22. The estimated relationship is robust to excluding banks with less than one percent insider holdings.

23. Results are not qualitatively different if the ratio of loan type to total loans is examined instead of the ratio of loan type to total assets.

24. We report region dummies in the case where the country is divided into four regions (North, South, East, West). We also experimented with eight regions (North, Northeast, Northwest, etc.) and twelve regions (corresponding to Federal Reserve districts), but the results are substantively the same.

these vectors, $\mathbf{X} = [\mathbf{O}|\mathbf{A}|\mathbf{N}|\mathbf{Yr}|\mathbf{Z}]$, the hypothesized relationship is of the form:

$$\mathbf{L_i} = \mathbf{X}'\beta + \mathbf{f}(\alpha) + \epsilon \qquad (\mathbf{15.4})$$

where $E(\epsilon\,|X, \alpha, L_1) = 0$ and where $\mathbf{W} = (\mathbf{L_1}, \mathbf{X}, \alpha)$ is identically distributed. The relationship, (15.4), consists of a parametric part, the term $\mathbf{X}'\beta$, and the nonparametric part, the function, $f(\alpha)$.[25]

Estimation of (15.4) and inference are complicated by the combination of the parametric and nonparametric components. Ordinary least squares regression of $\mathbf{L_i}$ on $\mathbf{X}$ would consistently and efficiently estimate β if $\mathrm{E}(\mathbf{X}\mathbf{f}(\alpha)) = 0$ which would occur, for example, if $\mathrm{E}(X) = 0$ and $\mathbf{X}$ were statistically independent of α. But, in our sample $\mathbf{X}$ and α are correlated since the largest banks tend to have smaller insider holding fractions. If we were interested primarily in β, then the bias in using OLS would be that of an omitted variable and there are a number of methods available to cope with this in a semiparametric context (see Heckman (1986, 1988), Robinson (1988), and Andrews (1990)). Our focus, however, is on the estimation of $f(\alpha)$ so we must take account of the parametric component in estimating the nonparametric part of the relationship. We use the semiparametric technique of Speckman (1988). Appendix 3 provides more detail on the estimation procedure.

15.4.3. Data

The data on loan portfolio shares are annual data from the *Call Reports* for the period 1984–1990. The annual data are not averaged so all right-hand side variables in the first step are measured annually except the outsider holdings (which are always for 1987 and 1988).[26] The parametric specification also includes year dummies to account for time affects. To avoid capturing situations where the incentives of managers and outside shareholders are aligned, we exclude observations where the ratio of equity capital to total assets is less than five percent (including these observations does not change the qualitative results).

15.4.4. Semiparametric Test Results

In Section 15.3 we established that during the 1980s consumer loans were relatively safe, while commercial real estate construction and development loans

25. The nonlinear relationship may be approximately quadratic (as in Proposition 3 above and McConnell and Servaes) or cubic (Morck, Shleifer, and Vishny) so in the parametric part of the relationship we include quadratic and cubic terms for total assets to ensure that such nonlinearities are not introduced spuriously by the parametric part of the estimation.

26. The shapes of the estimated functions are not affected by averaging data or varying window size, and are robust to shorter time periods.

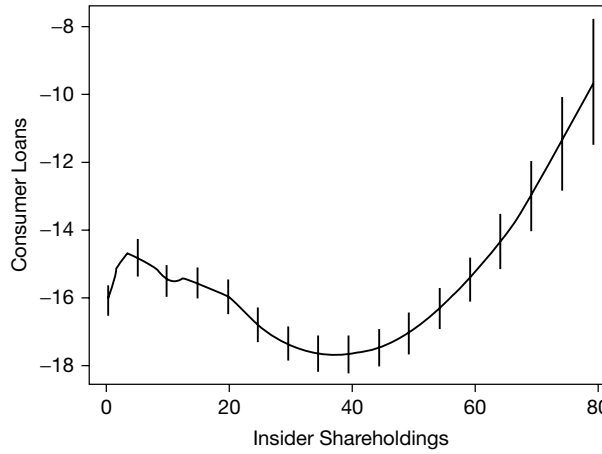


Figure 15.4 Results for the Nonparametric Component of the Semiparametric Regression of Consumer Loans Against Insider Shareholdings (α) and Control Variables, 1984–1990. The figure shows the estimated function, $\mathbf{f}(\alpha)$, which is the nonparametric component of: $\mathbf{L_i} = \mathbf{X}'\beta + \mathbf{f}(\alpha) + \epsilon$. This function was estimated using the Speckman (1988) method (described in Appendix 3). The dependent variable is consumer loans; the parametric component includes outside block shareholdings, log total asset, log total asset squared, regional dummies, and year dummies.

were risky. Figures 15.4 and 15.5 show the estimated nonlinear relationships between the loan shares of these two loan types and the fraction of equity held by insiders. Similarly, Figure 15.6 shows the estimated relationship for C&I loans, an intermediate category in terms of risk.[27] The vertical lines in the figures are 90 percent confidence intervals (see Cleveland and Devlin (1988)).

Figure 15.4 presents the estimated relationship between the fraction of consumer loans and insider holdings. At low levels of insider holdings, between zero and four percent, managers' interests move in the direction of outside shareholders, that is, they increasingly make relatively safe loans over this range. But, over the range from four to 40 percent, managers *reduce* their holdings of safe loans. Finally, for insider shares above 40 percent safe consumer lending increases, suggesting that at high levels of insider holdings interests become aligned; insiders basically become the owners. Thus, there appears to be a range where managers are entrenched; they take advantage of the power associated with their stockholding to make relatively few safe loans. At holdings of about 40 percent and above interests are aligned. The shape of the function in this case is similar to the U-shape imposed by McConnell and Servaes (1990).

27. The figures cut off the function at a level of insider holding of 80 percent for presentation purposes. No results are changed by this.

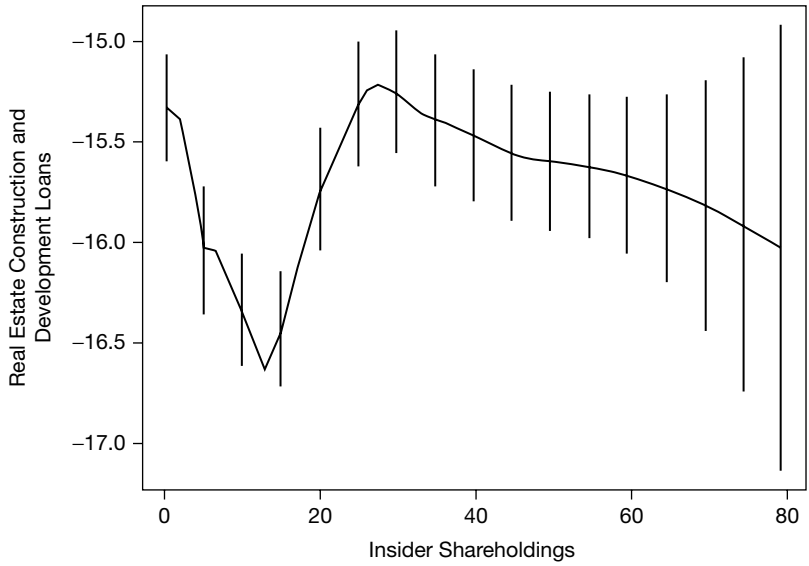


Figure 15.5 Results for the Nonparametric Component of the Semiparametric Regression of Real Estate Construction and Development Loans Against Insider Shareholdings (α) and Control Variables, 1984–1990. The figure shows the estimated function, $\mathbf{f}(\alpha)$, which is the nonparametric component of: $\mathbf{L_i} = \mathbf{X}'\beta + \mathbf{f}(\alpha) + \epsilon$. This function was estimated using the Speckman (1988) method (described in Appendix 3). The dependent variable is real estate construction and development loans; the parametric component includes outside block shareholdings, log total asset, log total asset squared, regional dummies, and year dummies.

Figure 15.5 shows the results for commercial real estate construction and development loans. Recall that these loans are the most risky. The pattern in Figure 15.5 is dramatically different from the pattern in Figure 15.4. In Figure 15.5 the pattern is a rotated s shape: over the range of insider holdings from zero to 15 percent, the share of the loan portfolio falls as insider ownership increases; from 15 to about 27 percent the function increases; it is flat from 27 to 50 percent and then declines, but the last decline is insignificant.[28] Confidence bands for higher fractions of insider holdings are very wide because we have few observations in that range. This pattern is similar to the pattern found by Morck, Shleifer, and Vishny (1988) who focused on Tobin's *q*.

Figure 15.6 presents the results for the intermediate category of commercial and industrial (C&I) loans. As expected the pattern is not as dramatic as for real estate construction and development loans and can be interpreted as falling in between the other two categories.

28. The pattern is very similar for the category of all commercial real estate loans.

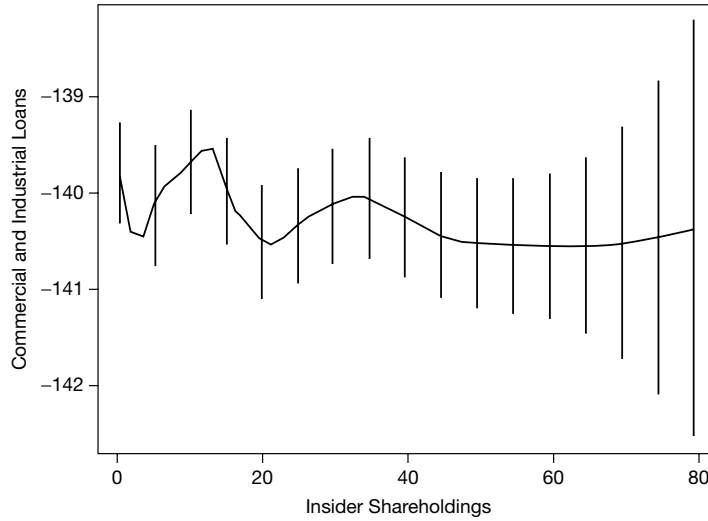


Figure 15.6 Results for the Nonparametric Component of the Semiparametric Regression of Commercial and Industrial Loans Against Insider Shareholdings (α) and Control Variables, 1984–1990. The figure shows the estimated function, $\mathbf{f}(\alpha)$, which is the nonparametric component of: $\mathbf{L_i} = \mathbf{X}'\beta + \mathbf{f}(\alpha) + \epsilon$. This function was estimated using the Speckman (1988) method (described in Appendix 3). The dependent variable is commercial and industrial loans; the parametric component includes outside block shareholdings, log total asset, log total asset squared, regional dummies, and year dummies.

With respect to the question of whether corporate control or moral hazard is better able to explain reality, the key question is the curvature of the above relationships.[29] The results are inconsistent with the moral hazard explanation of weakness in the U.S. banking system: risky loans are not made by managers with controlling interests; they make safe loans. At intermediate levels of stock holdings less than fifty percent, managers make relatively more risky, low-return, loans and fewer safe consumer loans. This is consistent with the view that these managers are entrenched. The results are also inconsistent with simple bad luck which we would not expect to be correlated with the fraction of stock held by insiders. We now turn to checking these initial results.

29. The level of the estimated curve is, fortunately, not important, since the intercept is not identified. This is because:

$$X'\beta + f(\alpha) = \left(\rho + X'\beta\right) + f(\alpha) - \rho$$

for all ρ. Since $f(\alpha)$ can always be redefined to be $f(\alpha) - \rho$, the intercept cannot be determined unless more structure is imposed. See Robinson (1988).

15.4.5. Results for the Parametric Specification

The advantage of the semiparametric estimation procedure is that it does not impose a functional form on $f(\alpha)$. For two reasons we also present the results of parametric estimation. First, the robustness of our results (in small sample) can be checked, using the parametric procedure. Parametric estimation is not robust in the sense of specification, since estimates are not consistent if the specification is incorrect, but, based on the semiparametric results, we can smooth the data more by imposing more structure. This can confirm our inferences in the sense that standard errors will be smaller (given that the parametric specification is consistent with the above results). Second, Proposition 3 predicts an inverse U-shaped pattern between insider holdings and riskier loans over the range where insiders are entrenched, and a U-shaped pattern for the relationship between insider holdings and relatively safe loans over the range where insiders are entrenched. By specifying a quadratic relationship between insider holdings and loan shares, restricting the sample to insider holdings between 10 and 80 percent, and including the variables from the first step into single estimation equation, we can test whether the predicted U-shaped patterns are present over the relevant range of insider holdings. Note that the quadratic specification which admits a U-shape or an inverse U-shape, and the limitation on the range of insider holdings, is consistent with the semiparametric results.

The results of these tests are shown in Table 15.3. Over the range of insider holdings of 10 to 80 percent the pattern for the relatively safe consumer loans is U-shaped, meaning that entrenched managers make fewer of these loans. On the other hand, the pattern for real estate construction and development loans is inversely U-shaped, that is, the entrenched managers make more of these risky loans. The pattern for commercial and industrial loans is U-shaped, but the coefficients are not significant. These results confirm our inferences from the previous procedure.

15.4.6. Further Results

A bank is a complicated set of activities and the mix of activities that different managers engage in, as a function of their opportunities and stock holdings, may well differ. For example, entrenched managers may engage in speculation on interest rates or trade foreign currencies, etc., but we have little data to determine the risk-return characteristics of these activities (compared to lending). Above, we examined the fairly specific predictions of the model about the lending choices of bank managers. We focus in this section on some additional possible implications of the model.

Table 15-3. Results of Quadratic Specification Tests on Various Loan Categories for Banks with Insider Holdings Between 10 and 80 Percent

The dependent variables in the regressions are the given loan category as a fraction of total assets. Inside and Inside^2 are insider ownership and insider ownership squared, in percentage points. Outside is the percentage of outside blockholder ownership. Log(TA) and Log(TA)^2 are log total assets and log total assets squared. The regional dummies, North, Midwest, South, and West, equal 1 if the bank is in the given region, and 0 otherwise. The year dummies, 1985 dummy–1990 dummy, are 1 if the observation is from that year and 0 otherwise. Each regression has 1212 observations. *t*-statistics are in parentheses.

	Dependent Variables		
	Consumer Loans	**Real Estate Constr. and Development Loans**	**Commercial and Industrial Loans**
Intercept	61.46	–65.58	47.69
	(1.85)	(3.39)	(1.07)
Inside	–0.33	0.12	–0.08
	(7.52)	(4.72)	(1.28)
Inside^2	0.005	–0.001	0.001
	(8.00)	(4.17)	(0.96)
Outside	–0.02	0.001	0.02
	(1.20)	(0.10)	(0.98)
Log(TA)	–10.72	13.64	–7.69
	(1.44)	(3.14)	(0.77)
Log(TA)^2	0.61	–0.71	0.52
	(1.46)	(2.92)	(0.93)
North	5.50	1.86	–3.30
	(3.36)	(1.95)	(1.50)
Midwest	3.39	0.08	–3.40
	(2.07)	(0.09)	(1.55)
South	2.90	2.32	–3.13
	(1.77)	(2.41)	(1.41)
West	1.91	5.61	4.01
	(1.13)	(5.70)	(1.77)
1985 dummy	–0.27	0.04	0.28
	(0.44)	(0.10)	(0.33)
1986 dummy	–0.97	0.22	–0.12
	(1.57)	(0.60)	(0.14)
1987 dummy	–1.48	0.59	–0.55
	(2.42)	(1.66)	(0.67)
1988 dummy	–1.73	0.80	–1.15
	(2.79)	(2.20)	(1.37)
1989 dummy	–2.08	0.54	–2.04
	(3.29)	(1.47)	(2.40)
1990 dummy	–2.70	0.38	–3.14
	(4.14)	(1.01)	(3.58)
Adjusted R^2	0.121	0.211	0.098

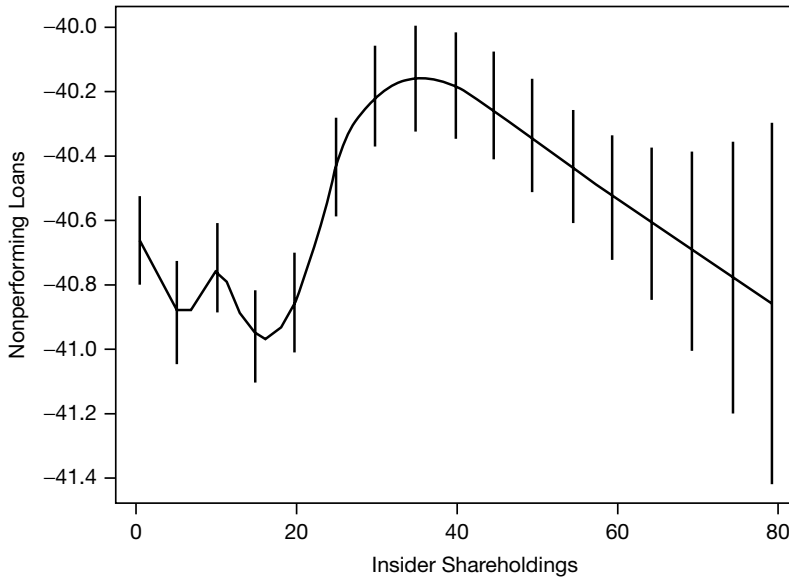


Figure 15.7 Results for the Nonparametric Component of the Semiparametric Regression of Nonperforming Loans Against Insider Shareholdings (α) and Control Variables, 1984–1990. The figure shows the estimated function, $\mathbf{f}(\alpha)$, which is the nonparametric component of: $\mathbf{L_i} = \mathbf{X}'\beta + \mathbf{f}(\alpha) + \epsilon$. This function was estimated using the Speckman (1988) method (described in Appendix 3). The dependent variable is nonperforming loans; the parametric component includes outside block shareholdings, log total asset, log total asset squared, regional dummies, and year dummies.

If entrenched bank managers make risky, low return loans, then we would expect them to suffer greater losses than other managers. Figure 15.7 shows the semiparametric estimate of the relationship between insider holdings and the ratio of nonperforming loans to total loans (controlling for other factors as we did earlier). Overall, the pattern has the rotated s shape. But, consistent with the above results, the relationship is roughly inversely-U-shaped over the range 10 to 80 percent. That is, over that range, entrenched managers have higher losses. This is confirmed with the quadratic specification results shown in Table 15.4.

If the risk-taking propensities of managers vary depending on how much equity they own, then this should be apparent in choices other than asset selection. On the liability side of the balance sheet, managers can increase risk by adding leverage. Figure 15.8 is the semiparametric estimate of the (book) equity-to-total-asset ratio. (Recall that all the banks in our sample satisfy regulatory capital requirements.) Again, the high leverage banks are those with managers in the entrenched range, consistent with the results. The parametric results are shown in Table 15.4.

Finally, if the corporate control hypothesis is correct, then we would predict that, looking to the future, banks with entrenched management would be less

Table 15-4. Results of Quadratic Specification Tests on Various Financial Ratios for Banks with Insider Holdings Between 10 and 80 Percent

The dependent variables in the regressions are nonperforming loans as a fraction of total loans, the ratio of equity capital to total assets, and the return on assets, all expressed as percentages. Inside and Inside^2 are insider ownership and insider ownership squared, in percentage points. Outside is the percentage of outside blockholder ownership. Log(TA) and Log(TA)^2 are log total assets and log total assets squared. The regional dummies, North, Midwest, South, and West, equal 1 if the bank is in the given region, and 0 otherwise. The year dummies, 1985 dummy–1990 dummy, are 1 if the observation is from that year and 0 otherwise. The first two regression have 1,212 observations, the final regression has 1,174 observations, *t*-statistics are in parentheses.

	Dependent Variables		
	Nonperforming Loans Ratio	**Equity-to Assets Ratio**	**Return on Assets**
Intercept	−5.13	14.54	−3.83
	(0.42)	(1.94)	(1.17)
Inside	0.05	−0.05	−0.002
	(2.93)	(4.87)	(0.48)
Inside^2	−0.001	0.00	0.00003
	(2.74)	(5.09)	(0.59)
Outside	0.01	−0.001	−0.004
	(1.85)	(0.07)	(2.56)
Log(TA)	1.77	−0.60	1.07
	(0.65)	(0.36)	(1.45)
Log(TA)^2	−0.12	−0.02	−0.06
	(0.79)	(0.19)	(1.42)
North	−0.65	−0.27	0.12
	(1.08)	(0.74)	(0.71)
Midwest	−0.60	−0.46	0.01
	(1.01)	(1.24)	(0.04)
South	0.92	−0.05	−0.10
	(1.52)	(0.13)	(0.58)
West	0.64	−0.26	0.02
	(1.05)	(0.68)	(0.11)
1985 dummy	0.18	0.04	−0.01
	(0.79)	(0.32)	(0.16)
1986 dummy	0.23	0.15	−0.07
	(1.01)	(1.11)	(1.14)
1987 dummy	0.15	0.41	−0.09
	(0.68)	(3.00)	(1.56)
1988 dummy	0.07	0.40	−0.00
	(0.30)	(2.87)	(0.06)
1989 dummy	0.29	0.54	−0.02
	(1.24)	(3.76)	(0.36)
1990 dummy	0.77	0.64	−0.15
	(3.22)	(4.34)	(2.34)
Adjusted R^2	0.119	0.148	0.025

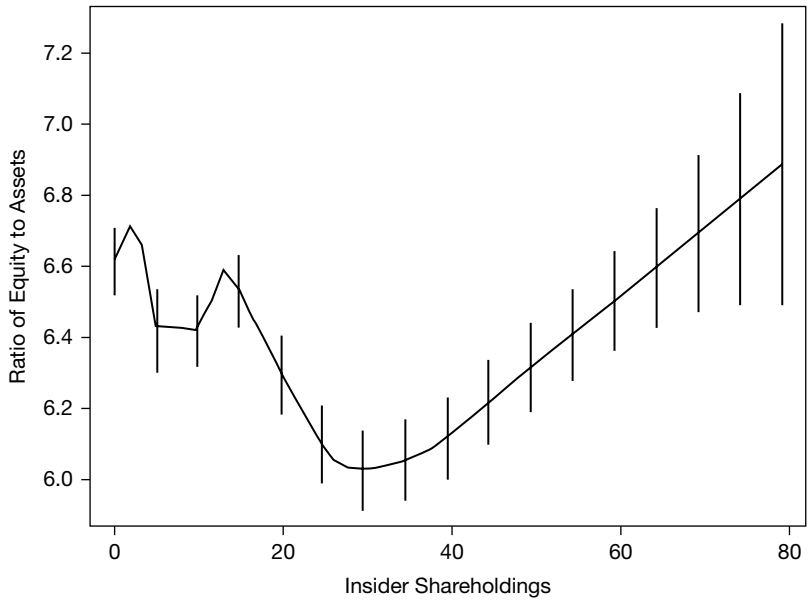


Figure 15.8 Results for the Nonparametric Component of the Semiparametric Regression of Equity-to-Assets Ratio Against Insider Shareholdings (α) and Control Variables, 1984–1990. The figure shows the estimated function, $\mathbf{f}(\alpha)$, which is the nonparametric component of: $\mathbf{L_i} = \mathbf{X}'\beta + \mathbf{f}(\alpha) + \epsilon$. This function was estimated using the Speckman (1988) method (described in Appendix 3). The dependent variable is the equity-to-assets ratio; the parametric component includes outside block shareholdings, log total asset, log total asset squared, regional dummies, and year dummies.

profitable. We can examine future rates of return to see if they reflect banks' equity ownership structure. We look at (book) return on assets (ROA) for the three years following our observation on managerial ownership. However, we find that there is no predictive power of the equity ownership structure for ROA (the figure is omitted, but Table 15.4 shows the parametric result). We also find (but do not show) similar results for (book) return on equity. We believe that survivorship bias against low-return and high-risk entrenched managers reduces our ability to find a significant relationship.

15.4.7. Summary

Overall, the empirical results confirm the pattern of lending behavior that the model of corporate control predicts. Notably, none of the results are what a moral hazard model would predict. The effect of moral hazard on bank decisions can vary. By relieving the need of insured depositors to monitor bank actions, deposit insurance makes it easier for banks to increase risk slightly. Deposit insurance can also lead banks with low charter values to "go for broke." The

moral hazard hypothesis should hold no matter the degree of moral hazard. If the effect of moral hazard is slight, however, it could potentially be overwhelmed by the effect of corporate control problems. Thus, while our results imply that corporate control problems are more important than moral hazard, we cannot conclude that deposit insurance has no effect on bank decisions. Our conclusion is that corporate control problems were empirically more important than moral hazard in explaining problems for large U.S. banks (which met regulatory capital requirements) during the 1980s. Moral hazard was not a significant problem.

15.5. CONCLUSION

Throughout the 1980s the U.S. banking industry systematically trended towards reduced profits and increased riskiness. The bank failure rate rose exponentially during the decade. It has been difficult to explain these trends. The previous literature tends to focus on the moral hazard hypothesis as an explanation, but evidence for this view has proved elusive. For example, Furlong (1988) finds that capital deficient bank holding companies in 1981 did not increase their risk over the next five years. McManus and Rosen (1991) do find a negative correlation between risk and return at banks, but only for banks above regulatory capital minimums. Banks with low capital levels appear to attempt to reduce risk, perhaps under regulatory pressure.

We propose an explanation for these trends based on corporate control problems in banking: outside equity holders do not make the lending decisions directly, but instead rely on managers. When bank managers receive private benefits of control, and outside shareholders can only imperfectly control them, managers will tend to take on excessive risk (relative to no agency costs) when the industry is unhealthy. This tendency is due to the incentives that managers face when the fraction of the bank they own is large enough for them to make outside discipline costly, but not so large as to cause their interests to be aligned with those of outsiders. This result contrasts with management behavior when the industry is healthy. In that case, the entrenched managers behave too conservatively.

We test the predictions of the model and find that, over the range of insider holdings where managers would tend to be entrenched, they make more risky loans (commercial real estate construction and development) and fewer relatively safe (consumer) loans. These results are consistent with the corporate control model, but contradict the pure moral hazard model (for banks with equity ownership structures over which the interests of managers and outside shareholders are not aligned). While we cannot rule out moral hazard, our findings suggest that corporate control problems have a bigger impact on bank risk-taking. (Mullins (1993) finds similar results: the relationship between insider

holdings and the standard deviation of stock returns in inversely U-shaped.) Since a joint hypothesis of the test was that the banking industry was unhealthy during the 1980s (i.e., characterized by declining investment opportunities), we have also provided evidence of this.

While our results suggest that corporate control problems are more important than moral hazard problems, our analysis is done for adequately-capitalized banks. If the value of bank equity is low enough, then the interests of inside and outside owners are aligned, so there are no corporate control problems of the sort we model. A reasonable interpretation of our results is that corporate control problems allow unprofitable banks to persist in making risky, low-return, loans. If, in the process, these banks lose enough equity value, then there may come a point at which inside and outside owners want to take excessive risk as the moral hazard hypothesis predicts. It may be accurate to say that, for large U.S. banks, corporate control problems have been the cause of the conditions of which moral hazard may be an accurate characterization.

The market for corporate control in banking is weaker than it is in markets for unregulated firms since regulation prevents nonbanks from taking over banks. The evidence on takeovers and takeover threats suggests that in the United States this is the main mechanism for disciplining managements (see Jensen and Ruback (1983)). Without the threat of nonbank takeovers it may be more difficult to induce bank managers to maximize shareholder value.[30] Consequently, the presence of agency costs suggests that the underlying trends that reduced profitability in the 1980s may persist, despite high bank earnings in the early 1990s. That banking is regulated does not appear to be a sufficient countervailing force.

"Banking" has traditionally corresponded to financing loans by issuing deposits. The combination of these activities has, historically, been the source of public policymakers' concerns. Firms called "banks" may eventually find other activities which are profitable, as Boyd and Gertler (1994) suggest, and transform themselves into viable entities which compete with other firms called "nonbanks," e.g., General Motors Acceptance Corporation. To the extent that chartered banks must transform themselves into nonbanks we say that "banking" is in decline. Whether chartered banks can survive by this transformation is not a question we consider. Our conclusions concern the difficulties that outside equityholders face during the transition period.

30. The importance of the takeover market in banking has been studied by James (1984) and James and Brickley (1987). Both studies examine the differences between two sets of banks: one set consists of states that prohibit corporate acquisitions of commercial banks, while the other set allows corporate acquisitions of banks. James (1984) finds that salary expenses, occupancy expense, and total employment are higher for banks in states which prohibit acquisitions. James and Brickley find that banks in states which allow acquisitions have more outside directors on their boards.

APPENDIX 1: EQUILIBRIUM WITH MANAGERIAL STOCK OWNERSHIP AND COSTLY FIRING

PROOF OF PROPOSITION 1: We compute the optimal response for managers given their beliefs about the firing rule used by outside owners. When firing rule (a) is used, a good manager is fired if and only if a risky loan portfolio is selected and gets a zero return. Thus, a good manager, maximizing expected return, makes risky loans if:

$$\begin{aligned}&\theta_G\left[(R-w)\,\alpha+w+(\theta_G R-w)\,\alpha+w\right]+(1-\theta_G)\left[-w\alpha+w+\alpha X\right]\\&\quad>(S_G-w)\,\alpha+w+(\theta_G R-w)\,\alpha+w.\end{aligned}\tag{15.5}$$

If the manager chooses a risky loan portfolio, the left-hand-side of (15.5), then with probability θ_G, the return is R. The manager gets the private benefits, w. To compute the return on the manager's stock, the private benefits, w, are deducted from the gross return so the manager's ownership share earns $(R-w)\,\alpha$. Since the loan return is R, the manager is allowed to continue to control the bank at date 2. Because the expected return on a good risky portfolio exceeds the expected return on a good safe portfolio, the manager chooses the risky portfolio at date 2 and expects to earn $(\theta_G R-w)\,\alpha+w$. If the return on the date 1 risky loan portfolio is zero, which occurs with probability $(1-\theta_G)$, then the manager is fired. Since the private benefits, w, are paid at date 1, as a shareholder, the manager must pay $w\alpha$, his share of the private benefits, to himself, and, as a manager, he receives private benefits of w. While he is fired, he remains a shareholder and receives αX, his share of the outsiders' best alternative at date 3.

If a safe loan portfolio is selected at date 1, the right-hand-side of (15.5), the manager receives his share of the return (net of the private benefits), $(S_G-w)\alpha$, plus the private benefits, w, at date 1. The return on his safe loan portfolio reveals him to be a good manager, so he is allowed to continue at date 2. At date 2 a good manager chooses a risky portfolio (because there is no distortion and it has a higher expected return than safe portfolio, by Assumption 1). Simplifying (15.5) shows that a manager chooses a risky loan portfolio if:

$$\Omega(\alpha)\equiv\left[\theta_G^2R-S_G+(X+w)\,(1-\theta_G)\right]\alpha-w\,(1-\theta_G)>0.\tag{15.6}$$

It is easy to see that $\Omega(0)=-w\,(1-\theta_G)<0$, so a good manager chooses a safe portfolio when he owns none of the bank. It also follows that:

$$\Omega(1)=\theta_G^2R-S_G+X\,(1-\theta_G)=\theta_G\,(\theta_G R-X)+(X-S_G)>0 \text{ for any } X,$$

so a good manager chooses a risky portfolio when he owns the bank and when he has committed to using firing strategy (a). More importantly, given the cost of firing a manager, we can show that there is a critical share α^* such that a good

manager chooses the safe portfolio for $\alpha < \alpha^*$ and the risky portfolio for $\alpha > \alpha^*$. Taking the derivative of $\Omega(\alpha)$ gives:

$$\begin{aligned}\Omega' &= \left[\theta_G^2 R - S_G + (X+w)(1-\theta_G)\right] + (1-\theta_G)\alpha X' \\ &= (1-\theta_G)\left(w/\alpha - \alpha c'\right) > 0 \quad \text{at} \quad \Omega = 0\end{aligned}$$

since $\left[\theta_G^2 R - S_G + (X+w)(1-\theta_G)\right] > 0$ whenever $\Omega = 0$ and $w > \alpha^2 c'$ by Assumption 4. Thus, since the function Ω is continuous, we know that there exists an α^* such that $\Omega(\alpha) \leq 0$ *if* $\alpha < \alpha^*$ and $\Omega(\alpha) \geq 0$ *if* $\alpha > \alpha^*$. In fact, we can solve for α^*:

$$\alpha^* = \text{Min}\left[\frac{w(1-\theta_G)}{\theta_G^2 R - S_G + (X+w)(1-\theta_G)} 1\right]. \tag{15.7}$$

Now consider the decisions of bad managers. Since firing rule (a) is assumed, bad managers choose risky portfolios if:

$$\begin{aligned}&\theta_B\left[(R-w)\alpha + w + (S_B - w)\alpha + w\right] + (1-\theta_B)\left[-w\alpha + w + \alpha X\right] \\ &\quad > (S_B - w)\alpha + w + \alpha X. \end{aligned} \tag{15.8}$$

Simplifying (15.8):

$$\Delta(\alpha) \equiv -\left[\theta_B(X+w-R) + (1-\theta_B)S_B\right]\alpha + w\theta_B > 0.$$

So, $\Delta(0) = w\theta_B > 0$ and $\Delta(1) = -\left[\theta_B(X-R) + (1-\theta_B)S_B\right]$ which can be either positive or negative since $X < R$ by Assumption 3. The derivative of Δ is:

$$\begin{aligned}\Delta' &= -\left[\theta_B(X+w-R) + (1-\theta_B)S_B\right] + \theta_B\alpha c' \\ &= -w\theta_B/\alpha + \theta_B\alpha < c' \quad \text{when} \quad \Delta = 0 \\ &= -\theta_B\left(w - \alpha^2 c'\right)/\alpha < 0 \ \text{ by } \ (A4).\end{aligned}$$

So, if $\Delta(1) > 0$, then a bad manager always chooses a risky portfolio, otherwise, since $\Delta' < 0$, there is a unique share of managerial ownership that is the dividing line between risky and safe portfolio choices:

$$\alpha^{**} = \text{Min}\left\langle \frac{\theta_B w}{\theta_B w - \theta_B(R-X) + (1-\theta_B)S_B}, 1 \right\rangle. \tag{15.9}$$

This completes the proof.

PROOF OF PROPOSITION 2: To prove Proposition 2, we need to solve the complete game between managers and outsiders. Given portfolio choices by managers, the expected return to an outsider (with one share) is $U_i(\psi, \phi)$ when outsiders choose firing rule $i \ln \{a, b, c\}$, good managers choose lending strategy $\psi \in \{\text{risky, safe}\}$, and bad managers choose lending strategy $\phi \in \ln \{\text{risky, safe}\}$.

When firing rule (a) is used, good managers choose a safe portfolio, and bad managers choose a risky portfolio, the expected return to outsiders is:

$$\begin{aligned} U_a\left(\text{safe, risky}\right) = {} & \gamma_{GG}\left[S_G + \theta_G R - 2w\right] \\ & + \gamma_G\left[\theta_G\left(1+\theta_G\right)R + \left(1-\theta_G\right)\left(X+w\right) - 2w\right] \\ & + \gamma_{BB}\left[\theta_B\left(R+S_B\right) + \left(1-\theta_B\right)\left(X+w\right) - 2w\right] \\ & + \gamma_B\left[\theta_B\left(1+\theta_B\right)R + \left(1-\theta_B\right)\left(X+w\right) - 2w\right]. \end{aligned}$$

A good (GG) manager chooses a safe portfolio at date 1. The return on the portfolio is S_G, of which shareholders get $S_G - w$, so the manager is allowed to continue control of the bank at date 2. Because the expected return on a good risky loan portfolio exceeds the expected return on a good safe portfolio, the good manager chooses a risky portfolio at date 2. The date 2 decision of the good manager offers the outsider an expected return of $(\theta_G R - w)$. A G manager chooses (per force) a risky portfolio at date 1. With probability θ_G, the return on the portfolio is R, so shareholders get $(R - w)$ after the manager take his private benefits. The manager is allowed to continue control of the bank at date 2, and chooses a risky portfolio, returning an expected $(\theta_G R - w)$ to outsiders. If the return on the risky portfolio selected at date 1 is zero, which occurs with probability $(1 - \theta_G)$, then the manager is fired. The private benefit is paid anyway and the outsider earns his expected opportunity cost X from the date 2 decision. A bad (BB) manager chooses a risky portfolio at date 1 and, if successful in avoiding being fired, chooses a safe portfolio at date 2. A B manager chooses a risky portfolio whenever he is in control.

The expected profit from firing rules (b) and (c) when good managers choose safe loans at date 1 and bad managers choose risky loans at date 1 can be similarly calculated. For firing rule (b),

$$\begin{aligned} U_b\left(\text{safe, risky}\right) = {} & \gamma_{GG}\left[S_G + \theta_G R - 2w\right] + \gamma_G\left[2\theta_G R - 2w\right] \\ & + \gamma_{BB}\left[\theta_B R + S_B - 2w\right] + \gamma_B\left[2\theta_B R - 2w\right]. \end{aligned}$$

For firing rule (c),

$$\begin{aligned} U_c\left(\text{safe, risky}\right) = {} & \gamma_{GG} S_G + \theta_G R - 2w\right] + \gamma_G\left[\theta_G R + X - w\right] \\ & + \gamma_{BB}\left[\theta_B R + X - w\right] + \gamma_B\left[\theta_B R + X - w\right]. \end{aligned}$$

Recall that the actions of the managers are taken as given in the above calculations. So, firing rule (a) is preferred by outsiders when good managers choose a safe portfolio and bad managers choose a risky portfolio if

$$U_a\left(\text{safe, risky}\right) > U_b\left(\text{safe, risky}\right) \qquad \textbf{(15.10)}$$

and

$$U_a\left(\text{safe, risky}\right) > U_c\left(\text{risky, safe}\right). \qquad \textbf{(15.11)}$$

(15.10) holds if:

$$\begin{aligned}\gamma_{GG}\left[S_G+\theta_G R-2w\right]+\gamma_G\left[\theta_G\left(1+\theta_G\right)R+\left(1-\theta_G\right)\left(X+w\right)-2w\right]\\ +\gamma_{BB}\left[\theta_B\left(R+S_B\right)+\left(1-\theta_B\right)\left(X+w\right)-2w\right]\\ +\gamma_B\left[\theta_B\left(1+\theta_B\right)R+\left(1-\theta_B\right)\left(X+w\right)-2w\right]\\ \geq\gamma_{GG}\left[S_G+\theta_G R-2w\right]+\gamma_G\left[2\theta_B R-2w\right]\\ +\gamma_{BB}\left[\theta_B R+S_B-2w\right]+\gamma_B\left[2\theta_B R-2w\right],\end{aligned}$$

which reduces to

$$\begin{aligned}\gamma_G\left(1-\theta_G\right)\left(X+w-\theta_G R\right)+\gamma_{BB}\left(1-\theta_B\right)\left(X+w-S_B\right)\\ +\gamma_B\left(1-\theta_B\right)\left(X+w-\theta_B R\right)\geq 0.\end{aligned}$$

Since $\theta_B R < S_B$ by Assumption 2, this is true if:

$$\left\langle\frac{\gamma_{BB}+\gamma_B}{\gamma_G}\right\rangle\left\langle\frac{X+w-S_B}{\theta_G R-X-w}\right\rangle\geq\frac{1-\theta_G}{1-\theta_B}.$$

(15.11) holds if

$$\begin{aligned}\gamma_{GG}\left[S_G+\theta_G R-2w\right]+\gamma_G\left[\theta_G\left(1+\theta_G\right)R+\left(1-\theta_G\right)\left(X+w\right)-2w\right]\\ +\gamma_{BB}\left[\theta_B\left(R+S_B\right)+\left(1-\theta_B\right)\left(X+w\right)-2w\right]\\ +\gamma_B\left[\theta_B\left(1+\theta_B\right)R+\left(1-\theta_B\right)\left(X+w\right)-2w\right]\\ \geq\gamma_{GG}\left[S_G+\theta_G R-2w\right]+\gamma_G\left[\theta_G R+X-w\right]\\ +\gamma_{BB}\left[\theta_B R+X-w\right]+\gamma_B\left[\theta_B R+X-w\right],\end{aligned}$$

which reduces to

$$\begin{aligned}\gamma_G\theta_G\left(\theta_G R-\left(X+w\right)\right)+\gamma_{BB}\theta_B\left(S_B-\left(X+w\right)\right)\\ +\gamma_B\theta_B\left(\theta_B R-\left(X+w\right)\right)\geq 0.\end{aligned}$$

Since $\theta_B R < S_B$ by Assumption 2, this is true if:

$$\frac{\theta_G}{\theta_B}\geq\left\langle\frac{\gamma_{BB}+\gamma_B}{\gamma_G}\right\rangle\left\langle\frac{X+w-\theta_B R}{\theta_G R-X+w}\right\rangle.$$

Similar calculations show $U_a(\text{risky, safe}) > U_b(\text{risky, safe})$ if

$$\left\langle\frac{\gamma_B}{\gamma_{GG}+\gamma_G}\right\rangle\left\langle\frac{X+w-S_B}{\theta_G R-X-w}\right\rangle\geq\frac{1-\theta_G}{1-\theta_B},$$

$U_a(\text{risky, safe}) > U_c(\text{risky, safe})$ if

$$\frac{\theta_G}{\theta_B}\geq\left\langle\frac{\gamma_B}{\gamma_{GG}+\gamma_G}\right\rangle\left\langle\frac{X+w-S_B}{\theta_G R-X-w}\right\rangle,$$

$U_a(\text{safe, safe}) > U_b(\text{safe, safe})$ if

$$\left\langle\frac{\gamma_B}{\gamma_G}\right\rangle\left\langle\frac{X+w-\theta_B R}{\theta_G R-X-w}\right\rangle\geq\frac{1-\theta_G}{1-\theta_B},$$

$U_a(\text{safe, safe}) > U_c(\text{safe, safe})$ if

$$\frac{\theta_G}{\theta_B} \geq \left\langle \frac{\gamma_B}{\gamma_G} \right\rangle \left\langle \frac{X + w - \theta_B R}{\theta_G R - X - w} \right\rangle,$$

$U_a(\text{risky, risky}) > U_b(\text{risky, risky})$ if

$$\left\langle \frac{\gamma_{BB} + \gamma_B}{\gamma_{GG} + \gamma_G} \right\rangle \left\langle \frac{X + w - S_B}{\theta_G R - X - w} \right\rangle \geq \frac{1 - \theta_G}{1 - \theta_B},$$

$U_a(\text{risky, risky}) > U_c(\text{risky, risky})$ if

$$\frac{\theta_G}{\theta_B} \geq \left\langle \frac{\gamma_{BB} + \gamma_B}{\gamma_{GG} + \gamma_G} \right\rangle \left\langle \frac{X + w - \theta_B R}{\theta_G R - X + w} \right\rangle.$$

It is clear from these inequalities that firing rule (a) dominates firing rule (b) for any strategies chosen by managers if (15.1) holds and that firing rule (a) dominates firing rule (c) for any strategies chosen by managers if (15.2) holds. This, along with Proposition 1 gives us the existence of a unique equilibrium. This completes the proof.

PROOF OF PROPOSITION 3: By (15.7) and (15.9),

$$\alpha^* - \alpha^{**} = \text{Min}\left[\frac{w(1-\theta_G)}{\theta_G^2 R - S_G + (X+w)(1-\theta_G)}, 1\right] - \text{Min}\left[\frac{\theta_B w}{\theta_B w - \theta_B (R - X) + (1-\theta_B) S_B}, 1\right].$$

When α^{**} and α^* are less than 1, then $\alpha^{**} > \alpha^*$ iff:

$$\begin{aligned}\theta_B \left(\theta_G^2 R - S_G + (1-\theta_G)(X+w)\right) - (1-\theta_G)\left((1-\theta_B) S_B - \theta_B R + \theta_B (X+x)\right) \\ = \theta_B \left(1 - \theta_G + \theta_G^2\right) R - \theta_B S_G - (1-\theta_G)(1-\theta_B) S_B \\ = \theta_B \left((1-\theta_G)^2 + \theta_G\right) R - \theta_B S_G - (1-\theta_G)(1-\theta_B) S_B \\ = \theta_B (1_G R - S_G) + (1-\theta_G)\left((1-\theta_G)\theta_B R - (1-\theta_B) S_B\right) > 0.\end{aligned}$$

The last line is the condition given in the proposition. Note that it is increasing in R and decreasing in S_G and S_B. The derivatives with respect to θ_G and θ_B are ambiguous. This completes the proof.

APPENDIX 2: EQUITY OWNERSHIP DATA

The data on the ownership structure of bank holding companies are constructed from 13D and 13G SEC filings as well as proxy statements, compiled by *Compact Disclosure. Compact Disclosure* was searched for data for the top 1274 bank holding companies. Usable data were found for 456 bank holding companies.

In many cases the holding company was not listed, presumably because it is not publicly held. In other cases, the data was not usable because it did not include the holdings of members of the board of directors. In a few cases the holdings added up to more than 100 percent of the outstanding stock; these cases are omitted.

The compilation lists all shareholders with at least five percent of the outstanding stock. To obtain the holdings of *outside* shareholders (with at least five percent), insider holdings are subtracted. Insider holdings are the amounts of stock held by officers and directors of the bank holding company. In addition, the following are counted as insiders: (1) director nominees; (2) stock in a holding company controlled pension fund or "ownership" plan; (3) stock held in trust for a director; (4) stock held by families of directors or officers; and (5) stock held by the bank's trust department, except when there are no other insiders. Excluded from the holdings of either insiders or outsiders is the stock of the parent company held by subsidiaries or stock of the bank which it holds itself. These two categories are treasury stock.

In the case of shares held by families of insiders, which are counted as inside holdings, the last name was used to identify families. For example, in the case of Jefferson Bankshares, Richard Crowell, Jr. is a director, but Richard Crowell, Sr. is not an officer or a director. Richard Crowell, Sr.'s stock is counted as an insider holding. Other examples are along the same lines. In general, the amount of inside holdings subtracted from the total outside holdings of those with at least five percent was added to the holdings of the remaining insider holdings.

The 13D and 13G other filing dates often differ from the dates of proxy filings. Sometimes dates were not provided. We used the most recent dates when dates were provided.

APPENDIX 3: SEMIPARAMETRIC ESTIMATION

To estimate (15.4) we follow Speckman (1988). Assume that the population regression function is a smooth function and that $\mathbf{X}$ and α are related via the regression model $\mathrm{E}(\mathbf{X}\,|\alpha) = \mathbf{g}(\alpha)$, i.e.,

$$\mathbf{X} = \mathbf{g}(\alpha) + \eta \qquad (\mathbf{15.12})$$

where η is a mean zero error term independent of α. The function $\mathrm{f}(\alpha)$ (see equation (15.4)) is estimated by assuming the existence of a smoother matrix, $\mathbf{K}$ for estimating the function $\mathrm{f}(\alpha)$ (we use locally weighted regression, as described below). Intuitively, $\mathbf{K}$ is the operator which, for each value of the nonparametric independent variable, calculates a value of the function at that point by attaching weights to neighboring points according to an assumed weighting function or density.

The smoother, **K**, cannot be applied directly to estimate the nonparametric part of the relationship, $f(\alpha)$, because of dependence on the parametric part, $\mathbf{X}^1\beta$. The basic approach is to purge each component of dependence on the other component, and then estimate the parametric part with OLS and the nonparametric part with a nonparametric estimator. Start by defining:

$$\mathbf{X}^* = (\mathbf{I} - \mathbf{K})\mathbf{X} \qquad \mathbf{L}_\mathbf{i}^* = (\mathbf{I} - \mathbf{K})\mathbf{L}_\mathbf{i}$$

which are the variables **X** and $\mathbf{L}_\mathbf{i}$; "adjusted" for dependence on α, via **K.** (**I** is the identity matrix.) Then β is estimated from partial residuals by:

$$\hat{\beta} = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{L}_\mathbf{i}^*.$$

And the estimate of the nonparametric component is given by:

$$\hat{\mathbf{f}} = \mathbf{K}\left(\mathbf{L}_\mathbf{i} - \mathbf{X}'\hat{\beta}\right).$$

Assuming (15.12), Speckman (1988) proves that: $n^{1/2}[\hat{\beta} - E(\hat{\beta})] \xrightarrow{D} N(0, \sigma^2\mathbf{V}^{-1})$ ($n^{-1}\eta'\eta \to \mathbf{V}$ where **V** is positive definite) and that the bias in estimating the nonparametric function, $f(\alpha)$, and its variance are negligible asymptotically. We now turn to a discussion of the choice of **K**.

We use locally weighted regression (see Cleveland and Devlin (1988), Müller (1987), Stute (1984), and Cleveland (1979)).[31] Local regression uses a weighted least squares estimate at each point using a neighborhood of the data points determined by choice of a window size or smoothing parameter, say *g*. The function $f(\alpha)$, at a point α_j (an element of α), $f(\alpha_j)$, is estimated by linear or quadratic weighted least squares. By varying the independent variable point, α_j, and recalculating the relevant neighborhood and weights at each point, the function can be traced out over its domain. Intuitively, the procedure is analogous to a moving average in time series analysis. Instead of averaging over time, however, the average is with respect to a neighborhood around each point (in cross-section).[32] Standard errors can be obtained following Cleveland and Devlin (1988).

31. The smoother matrix, **K,** may be linear or nonlinear (e.g., a low order polynomial) and possible methods include kernel, weighted regression, and spline procedures. (See Härdle (1990, 1991) and Muller (1988) for discussions.) The choice of locally weighted regression is due to the superior features of this method compared to kernel estimation. Local regression is more efficient that kernel methods and does not have "boundary effects" caused by the lack of a neighborhood on one side of data points near either end of the sample. These results are due to Fan (1992, 1993) and Stute (1984).

32. Note, however, that local regression is computationally burdensome even for samples of, say, $n = 200$ because at each point the sample must be sorted to find the q nearest neighbors. In time series the sorting is not an issue. In our case this issue is nontrivial because $n = 2{,}000$.

Local regression requires choice of a smoothing parameter, g. Thus, the estimate of $f(\alpha)$, say $\gamma_g(\alpha)$, depends on g and, therefore, the expected mean squared error also depends on g. The expected mean square error, S_g, is:

contribution of bias to the expected mean square error and V_g is the contribution of variance. Nonparametric estimators are biased (see Scott (1992)) when $\gamma_g(a)$ is a nearly unbiased estimate (which occurs when g is low, e.g., 0.2), then the expected value of B_g is nearly 0, but this depends on the choice of g. The difficulty is that choice of window size, g, trades-off variance of the estimator against bias.[33] There are a number of procedures for making the optimal choice of window size (which determines how smooth the estimated function is). However, our results do not change over a fairly broad range of window sizes.

REFERENCES

Andrews, Donald, 1990, Asymptotics for semiparametric econometric models: Estimation and testing, Discussion Paper No. 908R, Cowles Foundation.

Bagnani, Elizabeth, Nikolaos Milonas, Anthony Saunders, and Nickolaos Travlos, 1994, Managers, owners, and the pricing of risky debt: An empirical analysis, *Journal of Finance* 49, 453–478.

Barclay, Michael, and Clifford Holdemess, 1991, Negotiated block trades and corporate control, *Journal of Finance* 46, 861–878.

Bhattacharya, Sudipto, and Anjan Thakor, 1993, Contemporary banking theory, *Journal of Financial Intermediation* 3, 2–50.

Booth, James, 1993, FDIC improvement act and corporate governance of commercial banks, *Economic Review* 1, Federal Reserve Bank of San Francisco, 14–22.

Boyd, John, and Mark Gertler, 1994, Are banks dead? Or, are the reports greatly exaggerated?, Working paper, Federal Reserve Bank of Minneapolis.

Boyd, John, and Stanley Graham, 1991, Investigating the banking consolidation trend, *Quarterly Review* Spring, Federal Reserve Bank of Minneapolis, 3–15.

Boyd, John, and Edward Prescott, 1986, Financial intermediary-coalitions, *Journal of Economic Theory* 38, 211–32.

Cleveland, William, 1979, Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* 74, 829–836.

Cleveland, William, and Susan Devlin, 1988, Locally weighted regression: An approach to regression analysis by local fitting, *Journal of the American Statistical Association* 83, 596–610.

33. Bias and variance as $q \to \infty$, $n \to \infty$, and $g \to 0$ are given by:

$$\mathrm{E}\left[\gamma_g(\alpha)\right] - f(\alpha) \approx \frac{1}{24h(\alpha)^3}\left[\left(f''h + 2f'h'\right)(\alpha)\right]g^2 \quad \mathrm{Var}\left[\gamma_g(\alpha)\right] \approx \frac{\sigma^2(\alpha)}{g}$$

where $h(\alpha)$ is the marginal density of α. See Härdle (1991). Observe that the bias is increasing and the variance is decreasing in the smoothing parameter g.

Corrigan, E. Gerald, 1992, Rebuilding the financial strength of the U.S. banking system, *Quarterly Review,* Federal Reserve Bank of New York, 1–4.

Diamond, Douglas, 1984, Financial intermediation and delegated monitoring, *Review of Economic Studies* 51, 393–414.

Fama, Eugene, 1985, What's different about banks?, *Journal of Monetary Economics* 15, 5–29.

Fan, Jianqing, 1992, Design-adaptive nonparametric regression, *Journal of the American Statistical Association* 87, 998–1004.

Fan, Jianqing, 1993, Local linear regression smoothers and their minimax efficiencies, *The Annals of Statistics* 21, 196–216.

Federal Reserve Bulletin, Board of Governors of the Federal Reserve System, various issues.

Furlong, Frederick, 1988, Changes in bank risk-taking, Federal Reserve Bank of San Francisco, *Economic Review,* 45–55.

Gilson, Stuart, Kose John, and Larry Lang, 1990, Troubled debt restructurings: An empirical study of private reorganization of firms in default, *Journal of Financial Economics* 27, 315–354.

Gorton, Gary B., and Bruce Grundy, 1995, Corporate control, management stockholdings, and investment, Working paper, The Wharton School.

Gorton, Gary B., and George Pennacchi, 1995, Banks and loan sales: marketing nonmarketable assets, *Journal of Monetary Economics* 35, 389–412.

Gorton, Gary B., and Richard Rosen, 1992, Corporate control, portfolio choice, and the decline of banking, Working Paper #4247, National Bureau of Economic Research.

Härdle, Wolfgang, 1990, *Applied Nonparametric Regression* (Cambridge University Press, New York).

Härdle, Wolfgang, 1991, *Smoothing Techniques* (Springer-Verlag, New York).

Heckman, N., 1986, Spline smoothing in partially linear models, *Journal of the Royal Statistical Society* B 48, 244–248.

Heckman, N., 1988, Minimax estimates in a semiparametric model, *Journal of the American Statistical Association* 83, 1090–1096.

Holderness, Clifford, and Dennis Sheehan, 1988, The role of majority shareholders in publicly held corporations, *Journal of Financial Economics* 20, 317–346.

Hoshi, Takeo, Anil Kashyap, and David Scharfstein, 1990, The role of banks in reducing the costs of financial distress in Japan, *Journal of Financial Economics* 27, 67–88.

Houston, Joel, and Christopher James, 1993, An analysis of the determinants of managerial compensation in banking, Working paper, University of Florida.

James, Christopher, 1987, Some evidence on the uniqueness of bank loans, *Journal of Financial Economics* 19, 217–235.

James, Christopher, 1984, An analysis of the effect of state acquisition laws on managerial efficiency: The case of the bank holding company acquisitions, *Journal of Law and Economics* 27, 211–226.

James, Christopher, and James Brickley, 1987, The takeover market, corporate board composition, and ownership structure: The case of banking, *Journal of Law and Economics* 35, 161–180.

James, Christopher, and Peggy Weir, 1990, Borrowing relationships, intermediation, and the cost of issuing public securities, *Journal of Financial Economics* 28, 149–172.

Jensen, Michael, 1993, The modem industrial revolution, exit, and the failure of internal control systems, *Journal of Finance* 48, 831–880.

Jensen, Michael, and William Meckling, 1976, The theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics* 3, 305–360.

Jensen, Michael, and Richard Ruback, 1983, The market for corporate control: The scientific evidence, *Journal of Financial Economics* 11, 5–50.

Kane, Edward, 1992, The incentive incompatibility of government-sponsored deposit-insurance funds, in James Barth and Dan Brumbaugh, Eds.: *The Reform of Federal Deposit Insurance* (Harper Business, New York).

Keeley, Michael, 1990, Deposit insurance, risk, and market power in banking, *American Economic Review* 80, 1183–1200.

Lummer, Scott, and John McConnell, 1990, Further evidence on the bank lending process and the capital market response to bank loan agreements, *Journal of Financial Economics* 25, 99–122.

Marcus, Alan, 1990, Deregulation and bank financial policy, *Journal of Banking and Finance* 8, 557–565.

Marcus, Alan, and Israel Shaked, 1984, The valuation of FDIC deposit insurance using option-pricing estimates, *The Journal of Money, Credit and Banking* 16, 446–460.

McConnell, John, and Henri Servaes, 1990, Additional evidence on equity ownership and corporate value. *Journal of Financial Economics* 27, 595–612.

McManus, Douglas, and Richard Rosen, 1991, Risk and capitalization in banking, in *Rebuilding Banking,* Proceedings of a Conference on Bank Structure and Competition (Federal Reserve Bank of Chicago).

Merton, Robert, 1977, An analytic derivation of the cost of deposit insurance and loan guarantees, *Journal of Banking and Finance* 1, 3–11.

Mikkelson, Wayne, and Richard Ruback, 1985, An empirical analysis of the interfirm investment process, *Journal of Financial Economics* 14, 523–553.

Morck, Randall, Andrei Shleifer, and Robert Vishny, 1988, Management ownership and market valuation: An empirical analysis, *Journal of Financial Economics* 20, 293–316.

Müller, H. G., 1988, *Nonparametric Analysis of Longitudinal Data* (Springer-Verlag, Berlin).

Müller, H. G., 1987, Weighted local regression and kernel methods for nonparametric curve fitting, *Journal of the American Statistical Association* 82, 231–238.

Mullins, Helena, 1993, Risk-taking, managerial compensation and ownership structure: An empirical analysis, Working paper, University of British Columbia.

Pennacchi, George, 1987, The over- (or under-) pricing of deposit insurance, *Journal of Money, Credit and Banking* 19, 340–360.

Robinson, P. M., 1988, Root-n-consistent semiparametric regression, *Econometrica* 56, 931–954.

Ronn, Ehud, and Avinash Verma, 1986, Pricing risk-adjusted deposit insurance: An option-based model, *Journal of Finance* 41, 871–895.

Saunders, Anthony, Elizabeth Strock, and Nicholas Travlos, 1990, Ownership structure, deregulation, and bank risk-taking, *Journal of Finance* 45, 643–654.

Scott, David, 1992, *Multivariate Density Estimation* (John Wiley & Sons, New York).

Shleifer, Andrei, and Robert Vishny, 1986, Large shareholders and corporate control, *Journal of Political Economy* 94, 461–488.

Speckman, Paul, 1988, Kernel smoothing in partial linear models, *Journal of the Royal Statistical Society* B 50, 413–436.

Stulz, Rene, 1988, Managerial control of voting rights, *Journal of Financial Economics* 20, 25–54.

Stute, Winfried, 1984, Asymptotic normality of nearest neighbor regression function estimates, *Annals of Statistics* 12, 917–926.

Williams, Joseph, 1987, Perquisites, risk, and capital structure, *Journal of Finance* 42, 29–48.

Zeckhauser, Richard, and John Pound, 1990, Are large shareholders effective monitors? An investigation of share ownership and corporate performance, in Glenn Hubbard, Ed.: *Asymmetric Information, Corporate Finance, and Investment* (University of Chicago Press, Chicago).

16

Banks and Loan Sales Marketing Nonmarketable Assets

GARY B. GORTON* AND GEORGE G. PENNACCHI ■

16.1. INTRODUCTION

Historically, financial intermediaries have created loans that were not later sold. A reason for the illiquidity of loans is illustrated by the example of Penn Square, the bank that failed in 1982. According to the former director and chairman of the FDIC Irvine Sprague (1986, pp. 11–12):

> Penn Square was plunging other banks' money into the risky oil and gas business. Its mode of operation was to make large, high-priced but chancy loans to drillers and then to sell the loans, in whole or in part, to other banks while pocketing a fee for the service. Such loans are called 'participations' and are a common practice in banking. Penn Square, however, transformed the practice into a species of wheeling and dealing . . . The large participating banks were exposed, embarrassed, and threatened. Buying loan participations in enormous amounts were some of the country's leading and, supposedly, most sophisticated institutions . . . Their transactions with Penn Square violated all tenets of sound banking . . . They were content to rely on someone else's faulty and

* The data used in this paper were provided by a large bank which wishes to remain anonymous. We would like to thank the bank for the data and, particularly, the loan sales desk employees for their assistance with this study. Also, we are grateful for suggestions by seminar participants at the Federal Reserve Board, the Federal Reserve Banks of San Francisco and Cleveland, and Dartmouth College. We would also like to thank Mark Flannery, Stuart Greenbaum, Jonathan Karpoff, Deborah Lucas, Rabie Rafia, René Stulz, and an anonymous referee for providing helpful comments, and to the Geewax-Terker Research Program in Financial Instruments for research support.

> fragmentary loan documentation. Now they were exposed to massive and potentially fatal losses.

Subsequently, Seafirst of Seattle and Continental of Illinois, both major purchasers of Penn Square's loans, failed.

Recent theories of financial intermediation (e.g., Boyd and Prescott, 1986; Diamond, 1984) predict that purchasing loans would be treacherous. Banks provide borrowers with unique services in the form of (publicly unobserved) credit evaluation and monitoring activities. For a bank to have the incentive to provide an efficient level of these services, it is necessary that it hold (or retain the risk of) the loans that it creates. If loans were sold without recourse or guarantee to the buyer, then the bank would lack the incentive to produce an efficient level of credit information and monitoring since it would not receive the rewards from these activities. Ordinarily, loan buyers would recognize this lack of incentive and value the loan lower than otherwise. Therefore, the existence of financial intermediaries implies the creation of bank loans that banks should be unable to sell. The experience of Penn Square would seem to confirm the danger in buying loans and reinforce the presumption that bank loans are illiquid, which is the underlying rationale for much of bank regulation and Central Bank policy.[1]

The "participations" involved in Penn Square were secondary loan participations, more generally known as "commercial loan sales." These are contracts under which a bank sells a proportional (equity) claim to all or part of the cash flow from an individual loan to a third party buyer. The contract transfers no rights or obligations between the bank and the borrower, so the third-party buyer has no legal relationship with the bank's borrower. Furthermore, loan sales involve no type of recourse, credit enhancement, insurance, or guarantee because only then can the originating bank remove the loan from its balance sheet (according to regulatory accounting rules). In other words, the loan buyer has no recourse to the selling bank should a loan default occur.[2]

1. The nonmarketability of bank loans is often taken to imply that bank depositors have a difficult time valuing loans. It has been argued that such an information asymmetry between banks and outside investors is a precondition for banking panics. For example, Diamond and Dybvig (1983) assume that there is a cost to the bank of liquidating long-term investments. The cost is presumably motivated by the idea that such assets are nonmarketable. In Gorton (1985, 1986) banking panics are caused by depositor confusion over bank asset values.

2. The lack of recourse, guarantee, or credit enhancement sharply distinguishes secondary participations from other kinds of participation (novations and assignments). See Gorton and Haubrich (1989) for a discussion. Secondary participations are also unlike asset-backed securities in this respect. Not only are asset-backed securities typically credit-enhanced, but they are claims on the cash flows from a pool of loans, whereas a loan sale or secondary participation is a claim on the cash flow from a single loan.

Perhaps the problems inherent in selling loans, as exemplified by the Penn Square experience, explain why prior to the early 1980s loan sales never exceeded $20 billion annually and were confined to transactions within the bank correspondent network.[3] Important changes, however, occurred during the course of the 1980s when commercial and industrial loan sales grew tremendously, despite the practical experience and theoretical predictions that loan sales would be a "lemons" market. The amount of commercial and industrial loan sales outstanding, according to quarterly FDIC Call Reports, increased from approximately $26.7 billion in the second quarter of 1983 to a peak of $290.9 billion in the third quarter of 1989.[4] This growth was accompanied by a market that expanded beyond the confines of historical correspondent banking networks.[5] Also, the market developed from one where loans were primarily those of investment-grade firms to one where a majority of loans sold were non-investment-grade.

What explains the opening of the loan sales market?[6] A bank which needs to fund a new loan can: (1) fund the loan internally by issuing deposit liabilities having a cost defined as r_I, where r_I, includes any regulatory or agency costs associated with this source of financing, or (2) fund the loan by obtaining funds from a buyer of the loan, where this source of financing has a cost defined as r_{Is}. The fact that loan sales have not been observed in significant quantities for most of banking history suggests that internal funding costs were generally low compared to funding costs resulting from loan selling, as predicted by theories of financial intermediation. These theories suggest that the return the bank would have to promise a loan buyer, r_{Is}, would be higher than the bank's internal funding cost because, having sold the loan, the bank would lack the incentive to undertake costly credit risk analysis or monitoring. Realizing this, and the

3. According to American Bankers Association surveys, most loan sales in the correspondent network were due to overlines, i.e., instances where the originating bank exceeded its legal lending limit for an individual borrower.

4. Loan sales declined during the subsequent recession as the volume of new loans originated, especially loans financing mergers and acquisitions, declined. See Demsetz (1993/4), Demsetz (1994), Haubrich and Thomson (1993), and Cantor and Demsetz (1993).

5. Initially most loan purchasers were other banks (including a significant number of foreign banks), but nonbank firms accounted for about a quarter of loan purchases by the early 1990s (see Federal Reserve Board Senior Loan Officer Opinion surveys).

6. While some of the previous work on loan sales is discussed below, Berger and Udell (1993) provide a more complete summary. Previous empirical work, including Berger and Udell (1994), Carstrom and Samolyk (1993), Pavel and Phillis (1987), and Haubrich and Thomson (1993a, b), uses *Call Report* data to address questions concerning which banks are buyers and which banks are sellers of loans and also the variation of aggregate loan sales volume over time. Bernanke and Lown (1991) discuss loan sales and the "credit crunch."

resulting greater probability of the loan's default, loan buyers would demand a higher promised yield, r_{Is}, making loan sales relatively expensive.

Since we now observe significant quantities of loan sales, it appears that funding via loan selling is relatively inexpensive for some categories of loans originated by certain banks. This could be due to a rise in some banks' internal funding cost, r_I, and/or a decrease in the cost of funding loans via loan sales, r_{Is}. There seems to be little question that during the last fifteen years or so many banks' deposit funding costs have risen substantially. This period saw: (1) the lifting of interest rate ceilings on deposits (elimination of Regulation Q), (2) the development of interstate bank competition for deposits, and (3) increases in capital requirements that were binding constraints for many banks.[7] As shown in Pennacchi (1988) and Haubrich and Thomson (1993b), greater deposit market competition that leads to a rise in some banks' internal funding costs can result in an increase in aggregate loan sales, even if loan purchasers demand competitive rates of return on the loans that they purchase. This is because funds obtained from loan buyers, unlike deposit funds, avoid costs associated with required bank capital and required reserves. Banks facing competitive deposit markets will find that some loans can be profitably sold to certain smaller domestic banks or foreign banks that, due to local market power and/or regulation, have a relatively lower cost of deposit funds.

Could the rise in internal funding costs have led to loan sales that are nothing more than an implicit underwriting activity in which the originating bank provides no special credit evaluation or monitoring services? In other words, is loan selling simply a substitute for explicit commercial paper underwriting, a financing avenue available to mostly well-known investment-grade firms? This seems unlikely. If banks provided no special credit services, an explicit investment bank underwriting contract, which gives the investor a direct claim on the borrowing firm, would dominate a loan sales contract. Should the firm fail, the direct claim allows the holders legal rights that the indirect loan sale claim precludes. Only if banks continue to provide specialized credit services would loan selling be preferred over underwriting. In fact, loan selling does not appear to be a simple underwriting function involving no bank credit services. Most loans that have been sold were those of non-investment-grade firms. Indeed, for the money center bank studied later in this paper, the majority of its loan sales were claims on

7. It may also be the case that the internal funding costs are larger for particular categories of loans. Flannery (1989) argues that bank examination procedures create incentives for banks to hold only certain classes of loans, profitably selling the remainder. A significant fraction of loans sold during the 1980s were merger-related (see Federal Reserves Board Senior Loan Officer Opinion surveys). Loans to firms involved in highly levered transactions (HLT loans) faced particular regulatory pressure, suggesting that the costs of funding these loans internally was higher than for other categories of loans.

borrowers that did not have a commercial paper rating. Thus, a potential moral hazard problem, arising from a bank's lack of incentive to provide credit services when loans are sold, needs to be considered when discussing the cost of funding via loan sales.

While many banks' internal funding costs have likely increased, a decline in the cost of loan sales funding, r_{Is}, also may have occurred. This could help explain an expansion of the loan sales market. In Section 16.2 we present a model of loan sales that assumes that banks continue to provide unique credit services that are unobservable to loan buyers. We consider two possible contract features that could reduce the agency cost of selling loans. The first feature is the possibility of a bank offering an implicit guarantee on the value of a loan sold to the loan buyer. Regulation prevents banks from inserting explicit loan guarantees in loan sales contracts. There are, however, reasons to believe that an implicit guarantee may operate. Loan buyers are concerned with the lack of a secondary market where they could sell the participation should they need cash, so selling banks informally offer to buy back loans. The question is whether this process constitutes a form of insurance.[8] If a loan buyer expects the originating bank to buy back problem loans, a means of providing de facto loan guarantees would exist. The issue of implicit insurance has also been raised by regulators. For example, FDIC director Sprague (1986, p. 112) reported that the chairman of Penn Square "denied they had any hidden agreements to take back participated loans that went sour." Gorton and Pennacchi (1989), using loan sales yields averaged across a sample of banks, find very weak evidence of implicit bank guarantees on loan sales.

The other contract feature we examine concerns a bank's choice of selling only part of a loan. By retaining a portion of the loan, the bank could reduce agency problems since it continues to face a partial incentive to maintain the loan's value. The greater the portion of the loan held by the bank, the greater will be its incentive to evaluate and monitor the borrower. Notably, no participation contract requires that the bank selling the loan maintain a fraction, so this contract feature would also appear to be implicit and would need to be enforced by market, rather than legal, means. Simons (1993) considers the relation between the fraction of loan syndications held by the lead bank and credit quality.[9] We discuss Simons' results in comparison to our own later.

8. These statements are based on conversations with bankers and loan buyers. We were, unfortunately, unable to obtain data on the fraction of loan sales that were repurchased by the selling bank in our sample. Loan buyers and sellers report that loans are occasionally repurchased, but opinions varied as to whether the repurchase price amounted to (partial) insurance.

9. A loan syndication is not the same as a loan sale. In a loan sale the (legal) contractual relationship between the borrower and the bank is unaltered, but (part of) the cash flow promised by the borrower is sold to a third party with a new contract, the secondary participation. In a syndication a relationship between the borrower and the syndicate member is created from the beginning; in effect, there is no third party.

The model illustrates how these two contract features affect the equilibrium loan sales yield, r_{Is}, on a loan of a given credit class. It shows that if the loan is not fully guaranteed by the bank (implicitly), then the bank does not undertake the level of credit evaluation or monitoring that it would were it to hold the entire loan. The loan buyer recognizes this moral hazard and reduces the price it is willing to pay.

In loan sales made through the old correspondent banking network, the mechanism for enforcing implicit contracts may have involved the threat of loan buyers terminating other business relationships that they maintain with the originating bank. In today's environment, if a loan selling bank reneges on its implicit agreement to repurchase a loan or its commitment to retain a fraction of the loan, then potential buyers may not purchase the bank's loans in the future. Thus, failure to honor implicit agreements could lead to a loss of reputation and future profitable loan sales by the loan selling bank.[10]

In Section 16.3 we turn to empirical tests of the model. These tests use a unique data set of 872 loan sales. Unlike previous studies of loan sales, the data include deal-specific loan sales prices and the interest rates on the underlying loans. We use these data to test for the presence of the implicit contract features modeled in Section 16.2. Section 16.4 concludes.

16.2. A MODEL OF THE LOAN SALES MARKET

This section presents a model of the optimal contract between a bank and loan buyers. It considers a setting where the bank has an incentive to sell loans because of relatively high costs of internal funding.[11] Of course, banks may have other motives for loan sales, in particular, the desire to maintain a diversified loan portfolio. However, it seems hard to explain the dramatic 1980's rise in loan sales based solely on diversification, since this motive was likely to be present for most of banks' history. Various motivations for loan sales are discussed in Boyd and Smith (1989), James (1988), Pennacchi (1988), Benveniste and Berger (1987), Cumming (1987), Greenbaum and Thakor (1987), and

10. See Boot, Greenbaum, and Thakor (1993) for a model where reputation causes implicit financial guarantees to be fulfilled whenever the (bank) guarantor has sufficient financial capital.

11. High internal funding costs may be linked to a number of sources. Pennacchi (1988) shows that regulations, such as capital and reserve requirements, can add to the cost of competitively priced bank deposits to produce relatively high internal funding costs. James (1988) illustrates how a Myers (1977) type "underinvestment" problem can make deposit financing relatively costly when a bank has risky debt outstanding or is covered by fixed-premium deposit insurance.

Kareken (1987). See Berger and Udell (1993) for a review of the loan sales literature.

In the present model, a bank can improve the expected return on loans that it originates by evaluating (screening) loan applicants to identify better quality borrowers. However, as we explain below, the model can also be interpreted as one in which the bank provides alternative credit services by monitoring a borrower after originating a loan.[12] We adopt the standard assumption that the level of bank credit services is unobservable so that the bank and loan buyers cannot write contracts that are contingent on the level of these services.[13] Therefore, loan sales involve a moral hazard problem, namely, that the bank may not evaluate the credit of loan applicants at the most efficient level.

If a bank's diligence in screening loan applicants is unobservable, the consequent moral hazard problem can be mitigated by contractual features not directly concerned with the bank's effort. We consider the two features of the loan sale arrangement, discussed above, that could be contractually feasible: (i) an agreement by the bank to sell only a portion of the loan, retaining the remainder on its balance sheet, and (ii) a guarantee by the bank to repurchase the loan at a previously agreed upon price if the quality of the loan deteriorates. We interpret the second feature as equivalent to a (partial) guarantee against default on the loan sale. These two contract features can help mitigate the bank's moral hazard problem since the bank retains some of the risk of loan defaults and continues to face incentives to screen loan applicants.

16.2.1. Assumptions

The bank's problem is to maximize the expected profits from the sale of a particular loan.[14] The following assumptions are made about the loan's characteristics and possible contract features.

(A1) A bank loan requires one dollar of initial financing, and produces a stochastic return of x at the end of τ periods, where $x \in [0, L]$ and

12. Campbell and Kracaw (1980) and Boyd and Prescott (1986) explain the existence of financial intermediaries as providing efficient credit evaluation services. A monitoring role for intermediaries is modeled in Diamond (1984), Gorton and Haubrich (1987), and Gorton and Kahn (1994).

13. In recent years, the degree of asymmetric information between many borrowers and investors has likely declined, mitigating moral hazard problems in particular credit markets. However, complete elimination of asymmetric information between all potential borrowers and investors would leave banks with no role in credit intermediation. This seems extreme, so that we assume that a significant degree of asymmetric information continues to exist.

14. As shown in Pennacchi (1988), this problem is separable from the bank's choice of loan originations.

where L is the promised end-of-period repayment on the loan. The return, x, has a cumulative distribution function of $F(x, a)$, where a is the bank's level of credit evaluation. This distribution function satisfies

$$F(x, \lambda a + (1-\lambda)a') \leq \lambda F(x, a) + (1-\lambda)F(x, a') \quad \text{for all } a, a';\ \ \lambda \in (0, 1).$$

(A2) The bank has a constant returns to scale technology for evaluating the credit of loan applicants. The cost function is given by $c(a) = c \cdot a$.

(A3) The bank can sell a portion, b, of the return on a loan, where $b \in [0,\ 1]$, retaining the portion $(1 - b)$. Risk-neutral loan buyers require an expected rate of return on loans purchased of r_f. The bank finances its portion by issuing deposit and/or equity liabilities having the internal funding cost of r_I.

(A4) The bank has a policy of granting an implicit (partial) guarantee against the default of each loan that it sells. Let γ refer to the proportion of each loan sale that the bank promises to guarantee, where $\gamma \in [0,\ 1]$. The bank can fulfill this guarantee only if it is solvent at the time the loan matures. This future solvency of the bank is assumed to have probability p and to be uncorrelated with the return on the loan.

Assumptions (A1) and (A2) provide a rationale for a bank's services, improving the returns on loans by a costly credit evaluation of loan applicants.[15] We can view the bank as expending an unobserved level of credit screening service, a, in choosing to make a single loan to a particular applicant from a heterogeneous loan applicant pool. It is assumed that potential loan buyers know the risk distribution of the loan applicant pool, but they cannot observe the risk of an individual loan applicant within this risk class.[16] The distribution function of the loan that the bank ends up making from this risk class, $F(x,\ a)$, will be a function of its level of credit screening effort.

Due to the nature of the loan sales data that we subsequently analyze, our model focuses on a bank's credit evaluation services prior to originating loans. However, the model could be re-interpreted as one where the bank produces a variety of credit services. For example, virtually the same assumptions can characterize a situation where the bank provides costly monitoring services, such as in

15. These assumptions imply decreasing marginal profits from evaluating the credit of loan applicants.

16. For example, one particular risk class might be defined as all loan applicants that have no commercial paper rating. Within this (publicly observed) risk category, loan applicants could have varying degrees of (publicly unobserved) risk. Other risk classes might be those borrowers with A3, A2, Al, or A1 + commercial paper ratings.

Diamond (1984). The variable "a" can be viewed as the level of any bank service that increases the expected return on a loan.

Assumption (A3) constrains the form of the explicit loan sale contract to that of a proportional equity split between the bank and the loan buyer. This assumption is due to regulatory constraints that prevent other contract forms in selling commercial and industrial loans.[17] Assumption (A4) allows the bank to offer an implicit guarantee on the loans it sells. This level of guarantee is assumed to be the same for all loans that are sold.[18] The assumption that the bank's solvency and the return on a particular loan are uncorrelated can be justified if the loan is considered to be a small portion of the overall portfolio of assets (including off-balance sheet liabilities) held by the bank.

While assumption (A3) states that the bank is a price-taker in the market for loan sales (it must offer the expected rate of return of r_f to loan buyers), we place no restriction on the bank's market power in originating loans. In other words, banks may extract surplus from borrowing firms. We believe this is an important and realistic consideration, especially for borrowing firms that lack access to public security markets.[19] Hence our model, as well as our subsequent empirical work, does not assume that the yield on the loan paid by the borrowing firm reflects purely a risk premium or purely a monopoly rent.

16.2.2. The Bank's Problem

The optimal loan sales contract involves the bank's choice of credit screening effort, a, and the fraction of the loan to be sold, b, that maximizes its expected profits:

$$\max_{a,b} \int_0^L \left[(1-b)x - b_\gamma p(L-x)\right] \mathrm{d}F(x,a) - c(a) - e^{rf\tau} I, \qquad \textbf{(16.1)}$$

17. The constraints include restrictions on the form of a loan sale that enables a bank to remove the loan from its balance sheet, thereby avoiding reserve and capital requirements. Also, loan sales contracts must avoid the appearance of being "securities" in order to avoid securities laws. These issues are discussed by Gorton and Haubrich (1989).

18. The model can be extended to allow the bank to offer different implicit guarantees for each loan that it sells. This was done in an earlier version of this paper. Empirical results using this more complicated model are qualitatively similar.

19. Rajan (1992) presents a model where a bank's acquisition of firm-specific credit information gives it market power in making loans. Market power in bank lending is also consistent with empirical evidence regarding the incidence of reserve requirements analyzed in Fama (1985) and James (1987).

where

$$I = 1 - e_I^{-r_f \tau} \int_0^L \left[bx + b_\gamma p\,(L - x) \right] \mathrm{d}F\,(x, a)\,,$$

Subject to

(i) $\int_0^L \left(1 - b + b\gamma\, p\right)\, x\, \mathrm{d}F_a\,(x, a) = c'\,(a)\,,$

(ii) $b \leq 1.$

In problem (16.1), the first term in the bank's objective function is the expected return on the portion of the loan return held by the bank, minus the expected value of the implicit guarantee that the bank gives to the loan buyer, p is the probability that the bank is solvent (and can therefore honor its guarantee) when the loan matures in τ periods. I is the amount of internal (bank deposit and equity) funding that the bank must provide, at cost r_I, when a fraction b of the loan is sold. Constraint (i) is the incentive compatibility constraint. Hart and Holmstrom (1987) show that it can be written in this form when the distribution function, $F\,(x,\ a)$, satisfies the convexity-of-distribution-function condition given in (A1). Using the functional form $c(a) = c \cdot a$ and defining the expected return on the loan as

$$\overline{x}\,(a) = \int_0^L x\, \mathrm{d}F\,(x, a)\,,$$

the incentive compatibility constraint can be rewritten as

$$\overline{x}_a = \frac{c}{1 - b\left(1 - \gamma\, p\right)}, \tag{16.2}$$

where the subscript denotes partial differentiation. This constraint implies that when a bank sells a portion of the loan $(b > 0)$, and there is some probability of the bank failing $\left(p < 1\right)$ or the bank not fully guaranteeing the loan $(\gamma\ < 1)$, then the level of credit screening, a, is less than would be the case if the bank retained the entire loan $(b = 0)$ or credit screening was observable. In this latter case, credit screening could be set to its most efficient level, namely, that which satisfies[20]

$$\overline{x}_a = c. \tag{16.3}$$

The less-than-efficient level of credit screening that occurs when it is unobservable to loan buyers is the essence of the moral hazard problem that the bank attempts to minimize by other contractual arrangements. We now consider how the proportion of the loan that the bank sells, b, is optimally chosen to alleviate this problem.

20. Since the expected return on the loan is a concave function of the level of screening, a, comparing (16.2) and (16.3) implies a loss of efficiency when loans are sold.

16.2.3. Incentive Compatible Loan Sales

Problem (16.1) can be solved to jointly determine the equilibrium level of credit screening and the fraction of the loan to be sold. Define $\theta \equiv \exp\left[\left(r_I - r_f\right)\tau\right] - 1$ to be the excess cost of internal bank finance relative to financing at the risk-free rate.[21] Then the first-order conditions with respect to the bank's choices of b and a are

$$\{\theta\overline{x}(a) + \gamma p\theta[L - \overline{x}(a)] - \lambda(1 - \gamma p)\overline{x}_a - \mu\}b = 0, \tag{16.4}$$

$$\{[1 + b(1 - \gamma p)\theta]\overline{x}_a - c'(a) + \lambda[(1 - b(1 - \gamma p))\overline{x}_{aa} - c''(a)]\}a = 0, \tag{16.5}$$

where λ and μ are the Lagrange multipliers associated with constraints (i) and (ii), respectively. Assuming the interior solution ($a > 0$) and the functional form $c(a) = c \cdot a$, Eq. (16.2) can be substituted into Eq. (16.5) to eliminate c. The resulting expression can then be used to eliminate λ in Eq. (16.4). This produces the following equilibrium condition:

$$b = \frac{\theta\left[\overline{x}(a) + \gamma p\left(L - \overline{x}(a)\right)\right] - \mu}{\left(1 - \gamma p\right)\left[-\overline{x}_a^2/L\overline{x}_{aa}\right)\left(1 - \gamma p\right)(1 + \theta) + \theta\left[\overline{x}(a) + \gamma p\left(L - \overline{x}(a)\right)\right] - \mu}. \tag{16.6}$$

This condition will be the basis of our empirical tests. However, as currently written, Eq. (16.6) is difficult to interpret since it depends on the unobserved level and derivatives of the expected return on the loan, $\overline{x}(a)$. It can be simplified by replacing these unobserved expressions by observable variables or estimable parameters. First, we can substitute for $\overline{x}(a)$ by noting that it is directly related to the promised yield on the loan sold and the fraction of the loan guaranteed.

When a portion, b, of the loan is sold, the continuously compounded promised yield on the loan sale, r_{ls}, is defined by

$$r_{ls} = \frac{1}{\tau} ln\left(\frac{Lb}{1 - I}\right), \tag{16.7}$$

where $1 - I$ is the amount a loan buyer pays in return for the promised payment Lb. Substituting for I from problem (16.1) into Eq. (16.7) and rearranging, we obtain

$$\overline{x}(a) = \frac{L\left(e^{-\left(r_{ls} - r_f\right)^{\tau}} - \gamma p\right)}{1 - \gamma p}. \tag{16.8}$$

Second, in order to evaluate the ratio $\overline{x}_a^2/\overline{x}_{aa}$, we need to make an explicit assumption regarding the effect of credit screening on a given loan's expected return. We choose a simple parametric form that is consistent with our earlier

21. Note that θ is positive whenever $r_I > r_f$.

assumption about the bank's credit screening technology, assumption (A1), and also possesses sensible implications:

$$\overline{x}(a) = L\left(1 - \alpha e^{-\beta a}\right). \tag{16.9}$$

This functional form implies that if no credit evaluation is done ($a = 0$), the expected return on the loan is $L(1-\alpha)$. As credit services increase, the expected return on the loan asymptotes at the rate β to the promised payment, L.[22] Given Eq. (16.9), we have

$$-\overline{x}_a^2/\overline{x}_{aa} = L\alpha e^{-\beta a} = L - \overline{x}(a). \tag{16.10}$$

This expression, as well as Eq. (16.8), can then be used to simplify Eq. (16.6) as follows:

$$\begin{aligned} b &= \frac{\theta e^{-(r_{ls}-r_f)\tau} - \mu/L}{(1-\gamma p)\left[1+\theta - e^{-(r_{ls}-r_f)\tau} - \mu/L\right]} \\ &= \frac{r_I - r_f - \mu/(\tau L)}{(1-\gamma p)\left[r_I - r_f + r_{ls} - r_f - \mu/(\tau L)\right]}. \end{aligned} \tag{16.11}$$

By simple differentiation of Eq. (16.11), it is straightforward to prove:[23]

PROPOSITION. *In equilibrium, a bank sells a greater proportion of loans: (i) the greater is the bank's internal cost of funding, $r_I - r_f$; (ii) the lower is the equilibrium loan sale premium, $r_{ls} - r_f$; and (iii) the greater is the bank's probability of solvency, p.*

16.2.4. Discussion of the Model

The implications of the model, as summarized by the previous proposition, are intuitive. Banks will sell larger proportions of loans if they face a greater excess internal funding cost, since this is the direct cost of funding the part of the loan that they retain. They will also sell a greater proportion of less risky loans, those for which the provision of bank credit services is less vital, and for which loan buyers demand, in equilibrium, a smaller default premium. In addition, since an implicit guarantee to buy back a problem loan substitutes for loan retention as a way for banks to commit to efficient credit services, the greater the quality of this guarantee (the higher the bank's solvency probability), the less the proportion of the loan that the bank needs to retain.

22. The parameters α and β are assumed to be positive and loan specific. The parameter α is also assumed to be less than unity. The parameter β is a measure of the marginal increase in expected return on the loan from additional credit services.

23. For an interior equilibrium, $0 < b < 1$, the Lagrange multiplier, μ, equals zero.

Our result that banks will optimally sell a smaller fraction of more risky loans is consistent with empirical findings on loan syndications by Simons (1993). While loan syndications differ from loan sales in that the original loan contract is between the borrower and each syndicate member, one could argue that the lead bank (agent) managing the syndication plays a dominant role in credit evaluation. Also, the lead bank typically recruits syndicate members after making the initial contact with the borrower. Simons (1993) analyzed 1991 Shared National Credit Program data that reported bank regulators' classifications of syndicated loans and found that lead banks held a larger proportion of syndicated loans that were subsequently criticized by bank regulators.[24]

The model also suggests that banks choose less-than-efficient levels of credit screening when portions of loans are sold and not fully guaranteed. To the extent that bank loans differ from bonds by the provision of bank credit screening (or monitoring), this means that bank loans are "less special" when they are sold. Another interpretation is that "bank relationships" are less important when loans are sold. Of the 872 loan sales that we study in Section 16.3, 538 were sales in which the borrowing firm had no commercial paper rating, suggesting that if there is a decline in the significance of bank relationships, it is not only affecting large firms. However, recent research on very small firms suggests that bank relationships continue to be important (see Petersen and Rajan, 1993, 1994; Berger and Udell, 1994).

16.3. TESTS OF THE MODEL

This section considers the empirical validity of the model given in the previous section. The data are introduced first and the statistical tests follow.

16.3.1. An Overview of the Data

The data analyzed in this paper are a sample of 872 individual loan sales done by a major money center bank during the period January 20, 1987 to September 1, 1988. The bank, which has requested anonymity, is one of the largest loan sellers. For each loan sale, we were given the yield, maturity, and dollar size of the original loan made to the borrowing firm, the borrowing firm's commercial paper rating (if any), the yield and maturity of the loan sale, the fraction of the loan sold, and LIBOR corresponding to the date and maturity of the loan

24. On average, lead banks held a 17.4% stake in loans that were subsequently classified as "pass," while for criticized loans, lead banks held average loan proportions of 18.0%, 29.4%, 30.5%, and 47.3% for the classifications "specially mentioned," "substandard," and "loss," respectively.

Table 16-1. DESCRIPTION OF LOAN SALES DATA; JANUARY 20, 1987 TO SEPTEMBER 1, 1988; 872 OBSERVATIONS

Variable	Mean	Std. Dev.	Minimum	Maximum
Loan maturity (days)	28.04	22.45	1	277
Loan sale maturity (days)	27.63	22.44	1	277
Fraction of loan sold	0.76	0.30	0.09	1.00
Loan rate (%)	7.53	0.61	6.25	9.18
Loan sale rate (%)	7.41	0.59	6.28	9.12
LIBOR rate (%)	7.29	0.57	6.19	8.75

SOURCE: Money Center Bank.

sale.[25] In order that the yield on the original loan and the yield on the loan sold be comparable and not unduly reflect changes in market interest rates over the time interval between loan origination and loan sale, we restricted the sample to those loan sales that occurred within three days of the loan origination. This totaled 872 loan sale observations, or 90.1% of the original observations.[26] Table 16.1 gives summary statistics for this sample. Note that the average difference between the yield on the loan and the yield on the loan sale is approximately 12 basis points.[27] This is quite close to the average spread of 13 basis points that was found for money center banks during the Federal Reserve Board's June 1987 Senior Loan Officer Survey of Bank Lending Practices.

Table 16.2 stratifies loan sales by maturity and commercial paper rating. For each commercial paper rating and maturity category, the table provides the average size of the loan sale, the number of observations, the fraction of total observations falling into that cell, and the fraction of the all observations with the same maturity falling into that cell. Notably, the largest categories of sales (by number, but also by dollar volume) are those with maturities of 6–15 days and "No Rating", and 16–30 days and "No Rating". These two categories account for almost 47% of all loan sales. The next largest category is 31–60 days and "No Rating," which accounts for 10% of the total. Thus, these three categories account for over half the total sales. This is consistent with the earlier observation that loan sales may not simply be a substitute for commercial paper.[28]

25. The identity of the borrowing firm was not given to us.

26. Of this subsample of 872 loan sales, 74.8% were sales made on the date of origination, 15.4% were sales made one day after origination, 4.1% were sales made two days after origination, and 5.7% were sales made three days after origination.

27. Buyers of commercial and industrial loans do not pay or receive any additional fees when purchasing loans. They simply receive the promised yield on the participation.

28. Notably, this bank made no loan sales with maturities greater than one year, and its average maturity was about 28 days. This is shorter than the mean maturity of approximately one year reported by all banks during this time period. See Gorton and Haubrich (1989). The likely

Table 16-2. SUMMARY OF THE DATA: LOAN SALES SIZE, RATING, AND MATURITY

Rating	**Maturity (days)**					
	0–5	**6–15**	**16–30**	**31–60**	**61–90**	**90+**
A1+						
Average size of loan sale ($millions)	5.0	5.0	25.0	28.3	41.2	0
Number of observations	1	1	1	9	3	0
% of all observations	0.1	0.1	0.1	1.0	0.3	0
% of observations of same maturity	4.8	0.3	0.3	4.9	8.1	0
A1						
Average size of loan sale ($millions)	28.8	25.8	29.1	35.6	0	8.2
Number of observations	8	34	27	20	0	3
% of all observations	0.9	3.9	3.1	2.3	0	0.3
% of observations of same maturity	38.1	11.6	8.6	10.8	0	13.6
A2						
Average size of loan sale ($millions)	15.8	13.6	12.9	20.4	21.6	19.2
Number of observations	3	41	73	64	18	9
% of all observations	0.3	4.7	8.4	7.4	2.1	1.0
% of observations of same maturity	14.3	14.0	23.2	34.6	48.6	40.9
A3						
Average size of loan sale ($millions)	0	11.7	15.9	18.8	20.0	0
Number of observations	0	3	8	4	1	0
% of all observations	0	0.3	0.9	0.5	0.1	0
% of observations of same maturity	0	1.0	2.5	2.2	2.7	0
No rating						
Average size of loan sale ($millions)	16.1	11.0	13.4	18.9	15.8	14.9
Number of observations	9	210	206	88	15	10
% of all observations	1.0	24.1	23.6	10.1	1.7	1.1
% of observations of same maturity	42.9	71.9	65.4	47.6	40.5	45.5

Table 16.3 summarizes data that relates the spread of the yield on the loan negotiated with the borrower over LIBOR and the spread of the yield on the

explanation for the shorter average maturity is that none of the loan sales in our sample involved merger-related financings, which tend to have maturities in the range of five years. Other banks sold significant amounts of merger-related loans during this time period. These loans were almost always priced at 250 basis points over LIBOR, and there were no loans of this type in our sample.

Table 16-3. SUMMARY OF THE DATA: YIELD SPREADS (IN BASIS POINTS) AND FRACTION OF LOAN SOLD

Rating	**Maturity (days)**					
	0–5	**6–15**	**16–30**	**31–60**	**61–90**	**90+**
A1+						
Loan yield—LIBOR spread	−6.5	0.0	−5.0	18.6	60.8	—
Loan sale yield—LIBOR spread	−0.5	−2.0	−7.0	0.4	−3.3	—
Average fraction sold	1	1	1	0.843	0.556	0
Number of observations	1	1	1	9	3	0
A1						
Loan yield—LIBOR spread	12.4	2.9	8.9	1.8	—	30.4
Loan sale yield—LIBOR spread	3.9	−1.4	−3.5	1.9	—	6.0
Average fraction sold	0.917	0.867	0.746	0.455	—	1
Number of observations	8	34	27	20	0	3
A2						
Loan yield—LIBOR spread	5.1	6.2	9.1	18.4	23.1	22.9
Loan sale yield—LIBOR spread	4.8	4.7	3.1	5.8	10.5	12.7
Average fraction sold	0.733	0.826	0.810	0.746	0.600	0.608
Number of observations	3	41	73	64	18	9
A3						
Loan yield—LIBOR spread	—	22.8	15.7	14.6	25.0	—
Loan sale yield—LIBOR spread	—	17.5	12.7	12.0	17.5	—
Average fraction sold	—	0.778	0.771	0.625	1	—
Number of observations	0	3	8	4	1	0
No rating						
Loan yield—LIBOR spread	35.2	31.4	31.1	26.7	30.3	51.0
Loan sale yield—LIBOR spread	8.3	18.0	15.8	16.1	17.1	17.7
Average fraction sold	0.889	0.784	0.738	0.703	0.707	0.750
Number of observations	9	210	206	88	15	10

loan sale over LIBOR to the maturity of the loan and the rating of the borrower. Also given is the average fraction of each type of loan that the originating bank sells. Casual observation of Table 16.3 suggests that spreads generally increase as the borrower's rating declines and, perhaps, as the loan maturity lengthens. Also, the fraction of the loan sold by the bank appears to decline with maturity, holding the rating constant. However, there does not appear to be much relationship between the fraction sold and the rating of the borrower, holding maturity constant.

16.3.2. Testing the Specific Functional Form

Our first empirical test focuses on the equilibrium condition given by Eq. (16.11). As a means of empirically implementing the model, we assume that the natural logarithm of the proportion of a loan sold equals the natural logarithm of the right-hand side of Eq. (16.11) plus a normally distributed error term. Our hope is that this error term can capture the influence of missing factors, assumed to be uncorrelated with the right-hand side of Eq. (16.11), that determine the proportion of each loan sold. Because the natural log of the fraction of the loan sold, b, has a range between minus infinity and zero, Eq. (16.11) with an appended error describes a Tobit model. Defining b_i^* as a latent variable for loan sale i, and b_i as the observed variable (fraction sold) for loan sale i, we have

$$\ln\left(b_i^*\right) = -\ln\left(1-\gamma p\right) + \ln\left[\frac{\theta e^{-\left(r_{ls}-r_f\right)\tau}}{1+\theta-e^{-\left(r_{ls}-r_f\right)\tau}}\right] + \eta_i$$

$$\equiv z_i(\gamma p, \theta, (r_{ls}-r_f)\tau) + \eta_i, \qquad \textbf{(16.12a)}$$

$$b_i = b_i^* \quad \text{if} \quad 0 \le b_i^* \le 1, \qquad \textbf{(16.12b)}$$

$$b_i = 1 \quad \text{if} \quad 1 \le b_i^*, \qquad \textbf{(16.12c)}$$

where $\eta_i, \sim N\left(m, \sigma^2\right)$. Since the fraction of the loan sold, b_i, can at most be one, so that $\ln(b_i,)$ can at most be zero, the Tobit model is *censored* at $b_i = 1$. Therefore, the likelihood function is given by

$$\prod_{b_i<1} \frac{1}{\sigma}\phi\left(\frac{\ln(b_i) - m - z_i}{\sigma}\right) \prod_{b_i=1} N\left(\frac{m+z_i}{\sigma}\right), \qquad \textbf{(16.13)}$$

where ϕ is the standard normal probability density function.[29]

29. For example, see Maddala (1983, Ch. 6).

Recall that $\theta = \exp\left[\left(r_I - r_f\right)\tau\right] - 1$, so that the right-hand side of (16.12a) is a function of r_I, the bank's cost of internal financing. If we assume that the bank faces binding capital and reserve requirements, then the value of r_I can be written as[30]

$$r_I = \frac{r_e/(1-t) + \zeta r_d}{1 + \zeta(1-\rho)}, \tag{16.14}$$

where r_e is the cost (yield equivalent) of equity finance, r_d is the cost of deposit finance t is the corporate tax rate, ς is the bank's maximum debt–equity capital ratio, and ρ is the required reserve ratio on deposits. Our empirical work assumes a corporate tax rate, t, of 34%. Also, since most money center banks were near their minimum capital–asset ratio of 6% when these loan sales were made, we assume $0.06 = 1/(1+\varsigma)$.

The bank's marginal cost of deposit funds is assumed to equal the LIBOR yield having the same maturity as the loan sale, a measure that was provided to us along with the loan sales data. Since LIBOR is a nearly risk-free market rate, we assume it is equivalent to the quantity r_f in our model. The bank's reserve requirement on deposits, ρ, is assumed to be 3%. This was the amount of reserves required on nonpersonal time deposits, such as large Certificates of Deposit, during the sample period. The bank's cost of equity funds, r_e, is probably the most difficult rate to recover. In our empirical work, we make alternative assumptions that it equals the risk-free rate, r_f, or a constant spread over the risk-free rate, where this spread or "bank equity premium" is assumed to be 0.07, approximately the average difference between the rate of return on S&P 500 stocks and Treasury bills.

Estimating Eq. (16.12) also requires that we specify the probability of the bank failing by the maturity date of the loan sale. We assume that this probability is zero. This seems like a reasonable assumption due to the short maturity of the loan sales and the "too big to fail" doctrine followed by bank regulators.[31] Given our previous assumption that the bank's partial guarantee, γ, is the same for each loan, then the term $-\ln\left(1-\gamma p\right) = -\ln(1-\gamma)$ is a constant. While this implies that $-\ln(1-\gamma)$ is indistinguishable from m, the mean of the error term η_i, a quite literal interpretation of the model that assumes $m = 0$ would imply that γ could be estimated.

Employing the above assumptions, the Tobit model in Eq. (16.12) was estimated in the following form:

30. See Pennacchi (1988) for the simple derivation.

31. The alternative of estimating the failure probability from data on the bank's stock price was taken in an earlier version of this paper. Using these estimated failure probabilities, which averaged less than 0.0005 and had a maximum value of 0.017, produced qualitatively similar results.

Table 16-4. TEST OF THE MODEL'S SPECIFIC FUNCTIONAL FORM; 872 OBSERVATIONS; DEPENDENT VARIABLE: LOG OF FRACTION OF LOAN SOLD

$$\ln\left(b_i^*\right) = a_0 + a_1 \ln\left[\frac{\theta e^{-(r_{ls}-r_f)\tau}}{1+\theta-e^{-(r_{ls}-r_f)\tau}}\right] + \eta_i$$

Parameters	Tobit Model Parameter Estimates Assuming Equity Premium (E.P.) of 0, 0.04, 0.07 (standard errors in parentheses) (1) E.P. = 0	(2) E.P. = 0.04	(3) E.P. = 0.07
a_0	0.8338	0.8349	0.8340
	(0.1061)	(0.1070)	(0.1073)
a_1	0.5989	1.0094	1.3035
	(0.2479)	(0.4320)	(0.5680)
Value of γ implied from $a_0 = -\ln(1-\gamma)$	0.5656	0.5661	0.5657
	(0.0462)	(0.0465)	(0.0465)
Standard error, σ	1.6850	1.6868	1.6875
	(0.1096)	(0.1097)	(0.1097)

NOTE: The assumed equity premium is used in computing the cost of bank internal finance, r_I, which is a component of θ.

$$\ln\left(b_i^*\right) = a_0 + a_1 \ln\left[\frac{\theta e^{-(r_{ls}-r_f)\tau}}{1+\theta-e^{-(r_{ls}-r_f)\tau}}\right] + \eta_i, \qquad (16.15)$$

with the model restrictions being $a_0 = -\ln(1-\gamma)$ and $a_1 = 1$. The results of estimating Eq. (16.15) are given in Table 16.4.

As shown in columns 1 and 3 of Table 16.4, the model was first estimated assuming a bank equity premium of either 0 or 0.07. In either case, the estimates of a_0 and a_1 were consistent with the theoretical model. The a_1 estimates were positive and significantly different from zero at the 5% confidence level, but not significantly different from their theoretical value of 1.0. Since the equity premiums of 0 and 0.07 led to estimates of a_1 that straddled its theoretical value of 1.0, we then estimated the model assuming an intermediate equity premium of 0.04 and produced a statistically significant estimate of $a_1 = 1.0094$, almost identical to its theoretical value. See column 2. Hence, the model appears to be consistent with the data for a reasonable range of equity premia.

Given the assumption that $a_0 = -\ln(1-\gamma p)$ and $p = 1$, our estimates for a_0 in Table 16.4 imply a statistically significant value for the bank's partial guarantee of $\gamma = 0.57$. However, we would emphasize that while this estimate for γ does not seem unreasonable, our test of the hypothesis that the bank provides a partial guarantee is very weak. The estimate of γ is likely to be highly dependent on the functional form specified for the bank's credit screening technology, as well as the assumption that the disturbance term mean, m, is equal to zero. Hence, we must conclude that while the data is not inconsistent with the bank's giving a partial guarantee, there is certainly no strong evidence for this practice.

16.3.3. Testing the General Predictions of the Model

While the data appear consistent with the model as given by Eq. (16.15), its specific functional form does not allow us to distinguish how the loan sale risk premium, $(r_{ls} - r_f)\,\tau$, and the excess cost of internal bank financing, $(r_I - r_f)\,\tau$, independently influence the proportion of the loan sold, b. In this section we consider the general predictions of the model as summarized by the proposition of Section 16.2. The proposition suggests a test of the following relation:

$$b_i^* = \alpha_0 + \alpha_1 + (r_{ls} - r_f)\tau + \alpha_2(r_I - r_f)\tau + \varepsilon_i, \qquad \textbf{(16.16)}$$

where α_1 should be negative and α_2 should be positive. Since the fraction of the loan sold, b, is constrained to lie between $0 < b \leq 1$, a linear Tobit estimation technique was used. We first estimated Eq. (16.16) with the bank's cost of internal financing, r_I, calculated as before, assuming either an equity premium of 0 or 0.07. The results are given in columns 1 and 2 of Table 16.5.

Table 16.5 indicates that the coefficient on the loan sale risk premium, $(r_{ls} - r_f)\,\tau$, is correctly signed and statistically significant, verifying the model's prediction that the bank retains a greater proportion of the loan (sells less of the loan) for a larger equilibrium loan sale premium. In contrast, the coefficient of the internal funding cost variable, $(r_I - r_f)\,\tau$, is statistically insignificant, whether an equity premium of 0 or 0.07 is assumed. This insignificance may be due to the insensitivity of loan sales contracts to short-term movements in this variable.[32]

As an alternative to measuring a bank's excess cost of internal financing based on regulatory costs, we considered additional proxies for this cost based on the theory developed in James (1988). Briefly, this theory considers a situation in which banks have risky debt outstanding or are covered by fixed-premium deposit insurance. A Myers (1977) "underinvestment" problem can arise if the bank internally finances a new low risk loan because the new loan will lower the overall asset risk of the bank leading to a transfer of value from bank shareholders to bank debtholders or the FDIC. From the shareholders' perspective, this loss of value can be interpreted as a cost associated with internally financing low risk loans which can be avoided by loan sales. In contrast to low risk loans, the theory predicts that internally financing higher risk loans will be less costly since little, if any, value will be transferred from shareholders to debtholders or the FDIC. Thus, a measure of the safety or credit quality of a loan would be a proxy for the cost of internally funding the loan.

32. Differences in the excess cost of internal financing appear to better explain contemporaneous loan sales activity for a cross-section of different banks rather than loan sales activity across short time periods at the same bank. Using *Call Report* data for a cross-section of banks, Haubrich and Thomson (1993) find a statistically significant relation between a bank's loan sales and its cost of internal financing.

Table 16-5. TEST OF THE MODEL'S GENERAL IMPLICATIONS; 872 OBSERVATIONS; DEPENDENT VARIABLE: FRACTION OF LOAN SOLD

$$b_i^* = \alpha_0 + \alpha_1 + (r_{ls} - r_f)\,\tau + \alpha_2\,(r_I - r_f)\,\tau + \varepsilon_1$$

	Tobit Model Parameter Estimates (standard errors in parentheses)			
Explanatory Variables	**(1)**	**(2)**	**(3)**	**(4)**
Constant	0.4555	0.4564	0.4546	(0.3839)
	(0.0103)	(0.0095)	(0.0097)	(0.0515)
$(r_{ls} - r_f)\,\tau$	−131.47	−124.78	−155.87	−133.70
	(52.72)	(50.87)	(61.40)	(51.81)
$(r_I - r_f)\,\tau$ with E.P. $= 0$	3.959			
	(19.11)			
$(r_I - r_f)\,\tau$ with E.P. $= 0.07$		−0.1571		
		(1.1799)		
$(r_I - r_f)\,\tau$ proxied by $-(r_I - r_f)\,\tau$			−26.40	
			(30.48)	
$(r_I - r_f)\,\tau$ proxied by commercial paper ratings:				
Dummy $=$ 1 if *A*1				0.06435
				(0.05328)
Dummy $=$ 1 if *A*2				0.06356
				(0.5568)
Dummy $=$ 1 if *A*3				0.00996
				(0.07355)
Dummy $=$ 1 if no rating				0.08048
				(0.05195)
Standard error, σ	0.1388	0.1388	0.1390	0.1377
	(0.0055)	(0.0055)	(0.0003)	(0.0054)

NOTE: The assumed equity premium (E.P.) is used in computing the cost of bank internal finance, r_I.

We then re-estimated Eq. (16.16) by trying two different proxies for $(r_I - r_f)\tau$: minus the premium on the loan made to the borrower, $-(r_L - r_f)\tau$, and a set of dummy variables indicating the borrower's commercial paper rating, if any. Columns 3 and 4 of Table 16.5 display the results. While in both cases the coefficient on the loan sale risk premium, $(r_{ls} - r_f)\tau$, continues to be correctly signed and statistically significant, the proxies for $(r_I - r_f)\tau$ are insignificant. Thus, none of our measures for the bank's cost of internal financing appear to be strongly supported by the data.

16.4. CONCLUDING REMARKS

To better understand the opening of the loan sales market, we analyzed a model of bank and loan buyer behavior in which implicit contract features made loan

sales incentive compatible. If the selling bank retained a fraction of the loan or it gave loan buyers an implicit guarantee against default, this could explain why market participants would buy loans (assuming these implicit contracts could be enforced). The money center bank loan sales data that we analyzed were generally consistent with the model. In particular, the model's prediction that a bank will retain a greater proportion of more risky loans, that is, those with a higher equilibrium loan sale yield, was strongly supported by our empirical tests. While the data did not rule out the possibility of the bank giving implicit guarantees against default, the low power of our tests implies that the presence of this contractual feature continues to be an open question. However, considering the empirical evidence as a whole suggests that certain types of loans may not be perfectly liquid. A loan selling bank must continue to convince loan buyers of its commitment to evaluate the credit of borrowers by maintaining a portion of the loan's risk.

The existence of well-functioning markets for bank assets, like those which appear to be developing, does not mean that intermediation per se is ending. All the explanations for loan sales considered above imply that banks still offer services for certain classes of borrowers that cannot be obtained in capital markets via the underwriting of public securities. The loan sales contracts mean, however, that it is no longer necessary for banks to hold all loans until maturity, risking their capital during the life of the asset created.

REFERENCES

Benveniste, L. and A. Berger, 1987, Securitization with recourse: An investment that offers uninsured bank depositors sequential claims, Journal of Banking and Finance 11, 403–24.

Berger, A. and G. Udell, 1994, Lines of credit and relationship lending in small firm finance, Board of Governors of the Federal Reserve System working paper (Federal Reserve, Washington, DC).

Berger, A. and G. Udell, 1993, Securitization, risk, and the liquidity problem in banking, in: M. Klausner and L. White, eds., Structural change in banking (Irwin Publishing, Homewood, IL).

Bernanke, B. and C. Lown, 1991, The credit crunch, Brookings Papers on Economic Activity 2, 205–39.

Board of Governors of the Federal Reserve System, Reports of condition and income: Senior loan officer opinion survey on bank lending practices (Federal Reserve, Washington, DC), various issues.

Boot, A., S. Greenbaum, and A. Thakor, 1993, Reputation and discretion in financial contracting, American Economic Review 83, 1165–83.

Boyd, J. and E. Prescott, 1986, Financial intermediary-coalitions, Journal of Economic Theory 38, 211–32.

Boyd, J. and B. Smith, 1989, Securitization and the efficient allocation of investment capital, Federal Reserve Bank of Minneapolis working paper 408 (Federal Reserve Bank of Minneapolis, Minneapolis, MN).
Campbell, T. and W. Kracaw, 1980, Information production, market signalling and the theory of financial intermediation, Journal of Finance XXV, 863–81.
Cantor, R. and R. Demsetz, 1993, Securitization, loan sales, and the credit slowdown, Quarterly Review of the Federal Reserve Bank of New York, Summer, 27–38.
Carlstrom, C. and K. Samolyk, 1994, Loan sales as a response to market-based capital constraints, Federal Reserve Bank of Cleveland working paper (Federal Reserve Bank of Cleveland, Cleveland, OH).
Cumming, C., 1987, The economics of securitization. Quarterly Review of the Federal Reserve Bank of New York, Autumn, 11–23.
Demsetz, R., 1994, Economic conditions, lending opportunities, and loan sales, Federal Reserve Bank of New York working paper (Federal Reserve Bank of New York, New York, NY).
Demsetz, R., 1993/4, Recent trends in commercial bank loan sales, Quarterly Review of the Federal Reserve Bank of New York, Winter, 75–78.
Diamond, D., 1984, Financial intermediation and delegated monitoring, Review of Economic Studies LI, 393–414.
Diamond, D. and P. Dybvig, 1983, Bank runs, deposit insurance and liquidity, Journal of Political Economy 9, 401–19.
Fama, E., 1985, What's different about banks?, Journal of Monetary Economics 15, 5–29.
Flannery, M., 1989, Capital regulation and insured banks' choice of individual loan default rates, Journal of Monetary Economics 24, 235–58.
Gorton, G., 1985, Bank suspension of convertibility, Journal of Monetary Economics 15, 177–194.
Gorton, G., 1989, Self-regulating bank coalitions, Wharton School working paper (University of Pennsylvania, Philadelphia).
Gorton, G. and J. Haubtich, 1987, Bank deregulation, credit markets and the control of capital, Carnegie-Rochester Conference Series on Public Policy 26, 189–234.
Gorton, G. and J. Haubrich, 1990, The loan sales market, in: G. Kaufman, ed., Research in financial services: Private and public policy, Vol. 2 (JAI Press, Greenwich, CT).
Gorton, G. and J. Kahn, 1994, The design of bank loan contracts, collateral, and renegotiation, Wharton School working paper (University of Pennsylvania, Philadelphia, PA).
Gorton, G. and G. Pennacchi, 1989, Are loan sales really off-balance sheet?, Journal of Accounting, Auditing, and Finance 4, 125–145.
Greenbaum, S. and A. Thakor, 1987, Bank funding modes: Securitization versus deposits. Journal of Banking and Finance 11, 379–402.
Hart, O. and B. Holmstrom, 1987, The theory of contracts, in: T. Bewley, ed., Advances in economic theory: Fifth world congress (Cambridge University Press, Cambridge).
Haubrich, J. and J. Thomson, 1993a, The evolving loan sales market, Economic commentary of the Federal Reserve Bank of Cleveland, July 15 (Federal Reserve Bank of Cleveland, Cleveland, OH).
Haubrich, J. and J. Thomson, 1993b, Loan sales, implicit contracts, and bank structure, in: FDICIA: An appraisal, Proceedings of a conference on bank structure and competition (Federal Reserve Bank of Chicago, Chicago, IL).

James, C., 1987, Some evidence on the uniqueness of bank loans, Journal of Financial Economics 19, 217–35.
James, C., 1988, The use of loan sales and standby letters of credit by commercial banks, Journal of Monetary Economics 22, 395–422.
Kareken, J., 1987, The emergence and regulation of contingent commitment banking. Journal of Banking and Finance 11, 359–77.
Maddala, G.S., 1983, Limited-dependent and qualitative variables in econometrics (Cambridge University Press, Cambridge).
Myers, S., 1977, Determinants of corporate borrowing, Journal of Financial Economics 5, 147–175.
Pavel, C. and D. Phillis, 1987, Why commercial banks sell loans: An empirical analysis, Economic perspectives of the Federal Reserve Bank of Chicago, July/August (Federal Reserve Bank of Chicago, Chicago, IL), 3–14.
Pennacchi, G., 1988, Loan sales and the cost of bank capital, Journal of Finance 43, 375–95.
Petersen, M. and R. Rajan, 1994, The benefits of firm-creditor relationships: Evidence from small business data, Journal of Finance 49, 3–38.
Petersen, M. and R. Rajan, 1993, The effect of credit market competition on firm–creditor relationships, University of Chicago working paper (University of Chicago, Chicago, IL).
Rajan, R., 1992, Insiders and outsiders: The choice between informed and arm's-length debt, Journal of Finance 47, 1367–1400.
Simons, K., 1993, Why do banks syndicate loans?, New England Economic Review of the Federal Reserve Bank of Boston, January/February (Federal Reserve Bank of Boston, Boston, MA), 45–52.
Sprague, I., 1986, Bailout: An insider's account of bank failures and rescues (Basic Books, New York, NY).

17

Special Purpose Vehicles and Securitization*

GARY B. GORTON AND NICHOLAS S. SOULELES

17.1. INTRODUCTION

This paper analyzes securitization and more generally "special purpose vehicles" (SPVs), which are now pervasive in corporate finance.[1] What is the source of value to organizing corporate activity using SPVs? We argue that SPVs exist in large part to reduce bankruptcy costs, and we find evidence consistent with this view using unique data on credit card securitizations. The way in which the reduction in costs is accomplished sheds some light on how bank risk should be assessed.

By financing the firm in pieces, some on-balance sheet and some off-balance sheet, control rights to the business decisions are separated from the financing decisions. The SPV sponsoring firm maintains control over the business decisions while the financing is done in SPVs that are passive; they cannot make business decisions. Furthermore, the SPVs are not subject to bankruptcy costs because they cannot in practice go bankrupt, as a matter of design. Bankruptcy

* Thanks to Moody's Investors Service, Sunita Ganapati of Lehman Brothers, and Andrew Silver of Moody's for assistance with data. Thanks to Charles Calomiris, Richard Cantor, Mark Carey, Darrell Duffie, Loretta Mester, Mitch Petersen, Jeremy Stein, Rene Stulz, Peter Tufano, and seminar participants at the Philadelphia Federal Reserve Bank, Moody's Investors Service, and the NBER Conference on the Risks of Financial Institutions for comments and suggestions. Souleles acknowledges financial support from the Rodney L. White Center for Financial Research, through the NYSE and Merrill Lynch Research Fellowships.

1. Below we present the evidence on use of special purpose vehicles in the cases where such data exist. As explained below, these are "qualified" special purpose vehicles. Data on other types of SPVs are not systematically collected.

is a process of transferring control rights over corporate assets. Securitization reduces the amount of assets that are subject to this expensive and lengthy process. We argue that the existence of SPVs depends on implicit contractual arrangements that avoid accounting and regulatory impediments to reducing bankruptcy costs. We develop a model of off-balance sheet financing and test the implications of the model.

An SPV, or a special purpose entity (SPE), is a legal entity created by a firm (known as the sponsor or originator) by transferring assets to the SPV, to carry out some specific purpose or circumscribed activity, or a series of such transactions. SPVs have no purpose other than the transaction(s) for which they were created, and they can make no substantive decisions; the rules governing them are set down in advance and carefully circumscribe their activities. Indeed, no one works at an SPV and it has no physical location.

The legal form for an SPV may be a limited partnership, a limited liability company, a trust, or a corporation.[2] Typically, off-balance sheet SPVs have the following characteristics:

- They are thinly capitalized.
- They have no independent management or employees.
- Their administrative functions are performed by a trustee who follows prespecified rules with regard to the receipt and distribution of cash; there are no other decisions.
- Assets held by the SPV are serviced via a servicing arrangement.
- They are structured so that they cannot become bankrupt, as a practical matter.

In short, SPVs are essentially robot firms that have no employees, make no substantive economic decisions, have no physical location, and cannot go bankrupt. Off-balance sheet financing arrangements can take the form of research and development limited partnerships, leasing transactions, or asset securitizations, to name the most prominent.[3] And less visible are tax arbitrage-related transactions. In this paper we address the question of why SPVs exist.

The existence of SPVs raises important issues for the theory of the firm: What is a firm and what are its boundaries? Does a "firm" include the SPVs that it sponsors? (From an accounting or tax point of view, this is the issue of consolidation.) What is the relationship between a sponsoring firm and its SPV? In what sense does the sponsor "control" the SPV? Are investors indifferent between

2. There are also a number of vehicles that owe their existence to special legislation. These include REMICs, FASITs, RICs, and REITs. In particular, their tax status is subject to specific tax code provisions. See Kramer (2003).

3. On research and development limited partnerships see, e.g., Shevlin (1987) and Beatty, Berger, and Magliolo (1995); on leasing see, e.g., Hodge (1996, 1998), and Weidner (2000). Securitization is discussed in detail below.

investing in SPV securities and the sponsor's securities? To make headway on these questions we first theoretically investigate the question of the existence of SPVs. Then we test some implications of the theory using unique data on credit card securitizations.

One argument for why SPVs are used is that sponsors may benefit from a lower cost of capital because sponsors can remove debt from the balance sheet, so balance sheet leverage is reduced. Enron, which created over 3,000 off-balance sheet SPVs, is the leading example of this (see Klee and Butler (2002)). But Enron was able to keep their off-balance sheet debt from being observed by investors, and so obtained a lower cost of capital. If market participants are aware of the off-balance sheet vehicles, and assuming that these vehicles truly satisfy the legal and accounting requirements to be off-balance sheet, then it is not immediately obvious how this lowers the cost of capital for the sponsor. In the context of operating leases Lim, Mann, and Mihov (2003) find that bond yields reflect off-balance sheet debt.[4]

The key issue concerns why otherwise equivalent debt issued by the SPV is priced or valued differently than on-balance sheet debt by investors. The difference between on- and off-balance sheet debt turns on the question of what is meant by the phrase used above "truly satisfy the . . . requirements to be off-balance sheet." In this paper we argue that "off-balance sheet" is not a completely accurate description of what is going on. The difficultly lies in the distinction between formal contracts (which subject to accounting and regulatory rules) and "relational" or "implicit" contracts. Relational contracts are arrangements that circumvent the difficulties of formally contracting (that is, entering into an arrangement that can be enforced by the legal system).[5]

While there are formal requirements, reviewed below, for determining the relationships between sponsors and their SPVs, including when the SPVs are not consolidated and when the SPVs' debts are off-balance sheet, this is not the whole story. There are other, implicit, contractual relations. The relational contract we focus on concerns sponsors' support of their SPVs in certain states of the world, and investors' reliance on this support even though sponsors are

4. There are other accounting motivations for setting up off-balance sheet SPVs. E.g., Shakespeare (2001, 2003) argues, in the context of securitization, that managers use the gains from securitization to meet earnings targets and analysts' earnings forecasts. This is based on the discretionary element of how the "gain on sale" is booked. Calomiris and Mason (2004) consider regulatory capital arbitrage as a motivation for securitization, but conclude in favor of the "efficient contracting view," by which they mean that "banks use securitization with recourse to permit them to set capital relative to risk in a manner consistent with market, rather than regulatory, capital requirements and to permit them to overcome problems of asymmetric information . . ." (p. 26).

5. On relational contracts in the context of the theory of the firm see Baker, Gibbons, and Murphy (2002) and the references cited therein.

not legally bound to support their SPVs—and in fact under accounting and regulatory rules are not supposed to provide support.

The possibility of this implicit support, "implicit recourse," or "moral recourse" has been noted by regulators, rating agencies, and academic researchers. U.S. bank regulators define "implicit recourse" or "moral recourse" as the "provision of credit support, beyond contractual obligations . . ." See Office of the Comptroller of the Currency (OCC), et al. (2002, p. 1). The OCC goes on to offer guidance on how bank examiners are to detect this problem. An example of the rating agency view is that of FitchIBCA (1999): "Although not legally required, issuers [sponsors] may feel compelled to support a securitization and absorb credit risk beyond the residual exposure. In effect, there is moral recourse since failure to support the securitization may impair future access to the capital markets" (p. 4). Gorton and Pennacchi (1989, 1995) first discussed the issue of implicit recourse in financial markets in the context of the bank loan sales market; they also provide some empirical evidence for its existence.

Nonetheless, there are many unanswered questions. Why are SPVs valuable? Are they equally valuable to all firms? Why do sponsors offer recourse? How is the implicit arrangement self-enforcing? The details of how the arrangement works and, in particular, how it is a source of value has never been explained. We show that the value of the relational contract, in terms of cost of capital for the sponsor, is related to the details of the legal and accounting structure, which we explain below. To briefly foreshadow the arguments to come, the key point is that SPVs cannot in practice go bankrupt. In the U.S. it is not possible to waive the right to have access to the government's bankruptcy procedure, but it is possible to structure an SPV so that there cannot be "an event of default" which would throw the SPV into bankruptcy. This means that debt issued by the SPV should not include a premium reflecting expected bankruptcy costs, as there never will be any such costs.[6] So, one benefit to sponsors is that the off-balance sheet debt should be cheaper, ceteris paribus. However, there are potential costs to off-balance sheet debt. One is the fixed cost of setting up the SPV. Another is that there is no tax advantage of off-balance sheet debt to the SPV sponsor. Depending on the structure of the SPV, the interest expense of off-balance sheet debt may not be tax deductible.

After reviewing the institutional detail, which is particularly important for this subject, we develop these ideas in the context of a simple model and then test some implications of the model using data on credit card securitizations. The model analysis unfolds in steps. First, we determine a benchmark corresponding to the value of the stand-alone entity, which issues debt to investors in the capital

6. However, as we discuss below, the debt may be repaid early due to early amortization. This is a kind of prepayment risk from the point of view of the investors.

markets. For concreteness we refer to this firm as a bank. The bank makes an effort choice to create assets of types that are unobservable to the outside investors. Step two considers the situation where the assets can be allocated between on- and off-balance sheet financing, but the allocation of the assets occurs *before* the quality of individual assets has been determined. From the point of view of investors in the SPV's debt, there is a moral hazard problem in that the bank may not make an effort to create high-value assets. The sponsoring bank's decision problem depends on bankruptcy costs, taxes, and other considerations. We provide conditions under which it is optimal for the sponsoring bank to use an SPV.

The third step allows the bank to allocate assets *after* it has determined the qualities of its individual assets. In other words, investors in the debt issued by the SPV face an additional problem. In addition to the moral hazard associated with the effort choice, there is an adverse selection problem with regard to which projects are allocated to the SPV. We call this problem the "strategic adverse selection problem." In the case without commitment, investors will not buy the debt of the SPV because they cannot overcome the strategic adverse selection problem. However, we show that if the sponsor can commit to subsidize the SPV in states of the world where the SPV's assets are low quality and the sponsor's on-balance sheet assets are high quality, then the SPV is viable. In particular, if the bank can commit to subsidize the SPV in certain states of the world, then the profitability of the bank is the same as it would be when projects were allocated between the bank and the SPV prior to their realizations, i.e., when there was no strategic adverse selection.

But how does the commitment happen? Sponsors cannot verifiably commit to state-contingent subsidies. Even if they could verifiably commit to such strategies, legal considerations would make this undesirable because the courts view such recourse as meaning that the assets were never sold to the SPV in the first place. In this case, the SPV is not "bankruptcy remote," meaning that creditors of the sponsoring firm could "claw back" the SPV's assets in a bankruptcy proceeding. As Klee and Butler (2002) write:

> The presence of recourse is the most important aspect of risk allocation because it suggests that the parties intended a loan and not a sale. If the parties had intended a sale, then the buyer would have retained the risk of default, not the seller. The greater the recourse the SPV has against the Originator, through for example chargebacks or adjustments to the purchase price, the more the transfer resembles a disguised loan rather than a sale. Courts differ on the weight they attach to the presence of recourse provisions. Some courts view the presence of such a provision as nearly conclusive of the parties' intent to create a security interest, while others view recourse as only one of a number of factors. (p. 52)

This means that, as a practical matter, the recourse must not be explicit, cannot be formalized, and must be subtle and rare.

The final step in the analysis is to show that in a repeated context it is possible to implement a form of commitment. This result is based on the familiar use of trigger strategies (e.g. Friedman (1971), and Green and Porter (1984)), which create an incentive for the sponsor to follow the implicit arrangement. Previous applications of such strategies involve settings of oligopolistic competition, where firms want to collude but cannot observe strategic price or quantity choices of rivals. Intertemporal incentives to collude are maintained via punishment periods triggered by deviations from the implicit collusive arrangements. Our application is quite different. Here firms sponsoring SPVs "collude" with the investors in the SPVs by agreeing to the state-contingent subsidization of the SPV—recourse that is prohibited by accounting and regulatory rules. In this sense SPVs are a kind of "regulatory arbitrage."

Two empirically testable implications follow from the theoretical analysis. First, because the value in using SPVs derives in large part from avoiding bankruptcy costs, riskier firms should be more likely to engage in off-balance sheet financing. Mills and Newberry (2004) find that riskier firms use more off-balance sheet debt. Also, see Moody's (1997 September, 1997 January).

Second, following Gorton and Pennacchi (1989, 1995), implicit recourse implies that investors in the debt of the SPV incorporate expectations about the risk of the sponsor. This is because the sponsor must exist in order to subsidize the SPV in some states of the world. As Moody's (1997) puts it: "Part of the reason for the favorable pricing of the [SPVs'] securities is the perception on the part of many investors that originators (i.e., the "sponsors" of the securitizations) will voluntarily support—beyond that for which they are contractually obligated—transactions in which asset performance deteriorates significantly in the future. Many originators have, in fact, taken such actions in the past" (p. 40).

We test these two implications using unique data on credit card securitizations. We focus on securitization, a key form of off-balance sheet financing, because of data availability. Credit cards are a particularly interesting asset class because they involve revolving credits that are repeatedly sold into SPVs. Moreover, they represent the largest category within non-mortgage securitizations.

We find that, even controlling for the quality of the underlying assets and other factors, investors do require significantly higher yields for credit card ABS issued by riskier sponsors, as measured by the sponsors' credit ratings. Also, riskier firms generally securitize more, *ceteris paribus*. These results are consistent with our model.

The paper proceeds as follows. In Section 17.2 we provide some background information on off-balance sheet vehicles generally. Then, in Section 17.3 we focus more narrowly on some of the details of how securitization vehicles in particular work. Section 17.4 presents and analyzes a model of off-balance sheet

financing. In Section 17.5 we explain and review the data sets used in the empirical work. The first hypothesis, concerning the existence of implicit recourse, is tested in Section 17.6. The second hypothesis, that riskier firms securitize more, is tested in Section 17.7. Finally, Section 17.8 concludes, and is followed by a mathematical Appendix.

17.2. BACKGROUND ON SPVs

In this section we briefly review some of the important institutional background for understanding SPVs and their relation to their sponsor.

17.2.1. Legal Form of the SPV

A special purpose vehicle or special purpose entity is a legal entity which has been set up for a specific, limited purpose by another entity, the sponsoring firm. An SPV can take the form of a corporation, trust, partnership, or a limited liability company. The SPV may be a subsidiary of the sponsoring firm, or it may be an "orphan" SPV, one that is not consolidated with the sponsoring firm for tax, accounting, or legal purposes (or may be consolidated for some purposes but not others).

Most commonly in securitization, the SPV takes the legal form of a trust. Traditionally, a trust is "a fiduciary relationship with respect to property, arising as a result of a manifestation of an intention to create that relationship and subjecting the person who holds title to the property [the trustee] to duties with it for the benefit of [third party beneficiaries]" (Restatement (Third) of Trusts). Often the SPV is a charitable or purpose trust. These traditional trusts have been transformed into a vehicle with a different economic substance than perhaps contemplated by the law. These transformed trusts, commercial trusts, are very different from the traditional trusts (see Schwarcz (2003b), Langbein (1997), and Sitkoff (2003)).

A purpose trust (called a STAR trust in the Cayman Islands) is a trust set up to fulfill specific purposes rather than for beneficiaries. A charitable trust has charities as the beneficiaries. For many transactions there are benefits if the SPV is domiciled offshore, usually in Bermuda, the Cayman Islands, or the British Virgin Islands.

17.2.2. Accounting

A key question for an SPV (from the point of view of SPV sponsors, if not economists) is whether the SPV is off-balance sheet or not with respect to some other entity. This is an accounting issue, which turns on the question of whether

the transfer of receivables from the sponsor to the SPV is treated as a sale or a loan for accounting purposes.[7] The requirements for the transfer to be treated as a sale, and hence receive off-balance sheet treatment, are set out in Financial Accounting Standard No. 140 (FAS 140), "Accounting for Transfers and Servicing of Financial Assets and Extinguishment of Liabilities," promulgated in September 2000.[8] FAS 140 essentially has two broad requirements for a "true sale." First, the SPV must be a "qualifying SPV," and second, the sponsor must surrender control of the receivables.

In response to Enron's demise, the Financial Accounting Standard Board (FASB) adopted FASB Interpretation No. 46 (FIN 46) (revised December 2003), "Consolidation of Variable Interest Entities, an Interpretation of Accounting Research Bulletin (ARB) No. 51," which has the aim of improving financial reporting and disclosure by companies with variable interest entities (VIEs).[9] Basically, FASB's view is that the then current accounting rules that determined whether an SPV should be consolidated were inadequate. Because FASB had difficulty defining an SPV, it created the VIE concept. FIN 46 sets forth a new measure of financial control, one based not on majority of voting interests, but instead on who holds the majority of the residual risk and obtains the majority of the benefits, or both—independent of voting power.

A "qualifying" SPV (QSPV) is an SPV that meets the requirements set forth in FAS 140, otherwise it is treated as a VIE in accordance with FIN 46. FIN 46 does not apply to QSPVs. To be a qualifying SPV means that the vehicle: (1) is "demonstrably distinct" from the sponsor; (2) is significantly limited in its permitted activities, and these activities are entirely specified by the legal documents defining its existence; (3) holds only "passive" receivables, that is there are no decisions to be made; and (4) has the right, if any, to sell or otherwise dispose of non-cash receivables only in "automatic response" to the occurrence of certain events. The term, "demonstrably distinct," means that the sponsor cannot have the ability to unilaterally dissolve the SPV, and that at least ten percent of the fair value (of its beneficial interests) must be held by unrelated third parties.

On the second requirement of FAS 140, the important aspect of "surrendering control" is that the sponsor cannot retain effective control over the transferred assets through an ability to unilaterally cause the SPV to return specific assets

7. If the conditions of a sale are met, then the transferor must recognize a gain or loss on the sale.

8. Prior to FAS 140 the issue was addressed by FAS 125. FAS 140 was intended to clarify several outstanding questions left ambiguous in FAS 125.

9. VIEs are defined by FASB to be entities that do not have sufficient equity to finance their activities without additional subordinated support. It also includes entities where the equity holders do not have voting or other rights to make decisions about the entity, are not effectively residual claimants, and do not have the right to expected residual returns.

(other than through a cleanup call or to some extent "removal of accounts provisions").

FAS 140 states that the sponsor need not include the debt of a qualifying SPV-subsidiary in the sponsor's consolidated financial statements.

A QSPV must be a separate and distinct legal entity, separate and distinct, that is, from the sponsor (the sponsor does not consolidate the SPV for accounting reasons). It must be an automaton in the sense that there are no substantive decisions for it to ever make, simply rules that must be followed; it must be bankruptcy remote, meaning that the bankruptcy of the sponsor has no implications for the SPV; and the SPV itself must (as a practical matter) never be able to become bankrupt.

17.2.3. Bankruptcy

An essential feature of an SPV is that it be bankruptcy remote. This means that should the sponsoring firm enter a bankruptcy procedure, the firm's creditors cannot seize the assets of the SPV. It also means that the SPV itself can never become legally bankrupt. The most straightforward way to achieve this would be for the SPV to waive its right to file a voluntary bankruptcy petition, but this is legally unenforceable (see Klee and Butler (2002), p. 33 ff.). The only way to completely eliminate the risk of either voluntary or involuntary bankruptcy is to create the SPV in a legal form that is ineligible to be a debtor under the U.S. Bankruptcy Code. The SPV can be structured to achieve this result. As described by Klee and Butler (2002): "The use of SPVs is simply a disguised form of bankruptcy waiver" (p. 34).

To make the SPV as bankruptcy remote as possible, its activities can be restricted. For instance it can be restricted from issuing debt beyond a stated limit. Standard and Poor's (2002) lists the following traditional characteristics for a bankruptcy remote SPV:

- Restrictions on objects, powers, and purposes
- Limitations on ability to incur indebtedness
- Restrictions or prohibitions on merger, consolidation, dissolution, liquidation, winding up, asset sales, transfers of equity interests, and amendments to the organizational documents relating to "separateness"
- Incorporation of separateness covenants restricting dealings with parents and affiliates
- "Non-petition" language (i.e., a covenant not to file the SPE into involuntary bankruptcy)
- Security interests over assets

- An independent director (or functional equivalent) whose consent is required for the filing of a voluntary bankruptcy petition

The SPV can also obtain agreements from its creditors that they will not file involuntary petitions for bankruptcy. Depending on the legal form of the SPV, it may require more structure to ensure effective bankruptcy remoteness. For example, if the SPV is a corporation, where the power to file a voluntary bankruptcy petition lies with the board of directors, then the charter or by-laws can be structured to require unanimity. Sometimes charters or by-laws have provisions that negate the board's discretion unless certain other criteria are met.

An involuntary bankruptcy occurs under certain circumstances (see Section 303(b) of the Bankruptcy Code). Chief among the criteria is non-payment of debts as they become due. Perhaps most important for securitization vehicles, shortfalls of cash leading to an inability to make promised coupon payments can lead to early amortization rather than an event of default on the debt. This is discussed further below.

There is also the risk that if the sponsor of the SPV goes bankrupt, the bankruptcy judge will recharacterize the "true sale" of assets to the SPV as a secured financing, which would bring the assets back onto the bankrupt sponsor's balance sheet. Or the court may consolidate the assets of the sponsor and the SPV. As a result of this risk, most structured financings have a two-tiered structure involving two SPVs. The sponsor often retains a residual interest in the SPV that provides a form of credit enhancement, but the residual interest may preclude a "true sale." Consequently, the residual interest is held by another SPV, not the sponsor. The "true sale" occurs with respect to this second vehicle. This is shown in Figure 17.1, which is taken from Moody's (August 30, 2002).

17.2.4. Taxes

There are two tax issues.[10] First, how is the SPV taxed? Second, what are the tax implications of the SPV's debt for the sponsoring firm? We briefly summarize the answers to these questions.

The first question is easier to answer. SPVs are usually structured to be tax neutral, that is, so that their profits are not taxed. The failure to achieve tax neutrality would usually result in taxes being imposed once on the income of the sponsor and once again on the distributions from the SPV. This "double tax" would most likely make SPVs unprofitable for the sponsor. There are a number

10. This subsection is based on Kramer (2003), Peaslee and Nirenberg (2001), and Humphreys and Kreistman (1995).

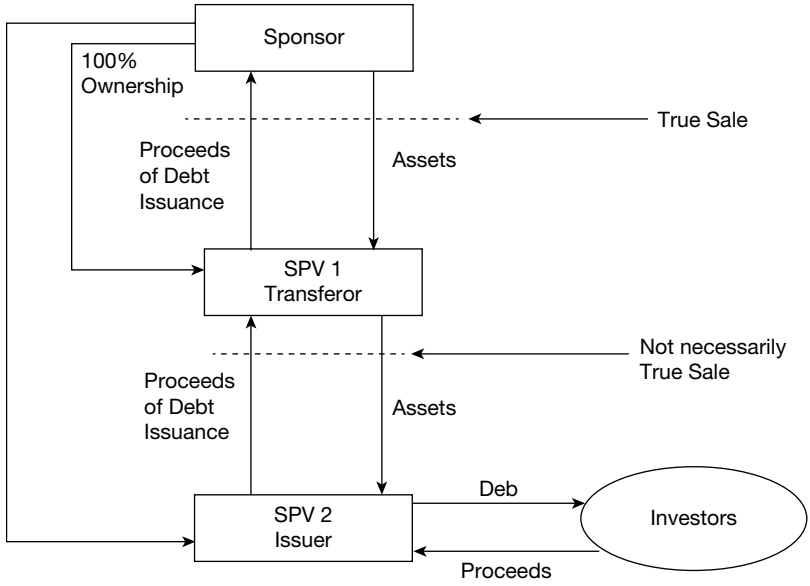


Figure 17.1 A Two-Tiered Bankruptcy Remote Structure
SOURCE: Moody's (August 30, 2002).

of ways to design an SPV to achieve tax neutrality. We briefly review some of them.

Many SPVs are incorporated in a tax haven jurisdiction, such as the Cayman Islands, where they are treated as "exempted companies." See Ashman (2000). An exempted company is not permitted to conduct business in, for example, the Cayman Islands, and in return is awarded a total tax holiday for twenty years, with the possibility of a ten-year extension. Because such entities are not organized or created in the U.S., they are not subject to U.S. federal income tax, except to the extent that their income arises from doing business in the U.S. However, the organizational documents for the SPV will limit it so that for purposes of the U.S. Internal Revenue Code of 1986, it can be construed as not being "engaged in U.S. trade or business."

An investment trust that issues pass-through certificates is tax neutral; that is, the trust is ignored for tax purposes—there is no taxation at the trust level—and the certificate owners are subject to tax. Pass-through certificates represent pro rata interests in the underlying pool. To maintain this tax-neutral tax status, it is important that the SPV not be reclassified as a corporation. To avoid such reclassification, the trustee must have no power to vary the investments in the asset pool, and its activities must be limited to conserving and protecting the assets for the benefit of the beneficiaries of the trust. See Kramer (2003).

More common than pass-through structures are pay-through structures. Pay-through bonds are issued by SPVs that are corporations or owner trusts. In these structures the SPVs issue bonds, but this requires that there be a party that holds

the residual risk, an equity holder. If the SPV is a corporation, then the pay-through bonds have minimal tax at the corporate level because the SPV's taxable income or loss is the difference between the yields on its assets and the coupons on its pay-through bonds. Typically these are matched as closely as possible.

The second question is more complicated. Some SPVs achieve off-balance sheet status for accounting purposes but not for tax purposes. Securitizations can fit into this category because they can be treated as secured financing for tax purposes.

17.2.5. Credit Enhancement

Because the SPV's business activities are constrained and its ability to incur debt is limited, it faces the risk of a shortfall of cash below what it is obligated to pay investors. This chance is minimized via credit enhancement. The most important form of credit enhancement occurs via tranching of the risk of loss due to default of the underlying borrowers. Tranching takes the form of a capital structure for the SPV, with some senior rated tranches sold to investors in the capital markets (called A notes and B notes), a junior security (called a C note) which is typically privately placed, and various forms of equity-like claims. Credit enhancement takes a variety of other forms as well, including over-collateralization, securities backed by a letter of credit, or a surety bond, or a tranche may be guaranteed by a monoline insurance company. There may also be internal reserve funds that build-up and diminish based on various criteria. We review this in more detail below with respect to credit card securitization in particular.

17.2.6. The Use of Off-Balance Sheet Financing

Off-balance sheet financing is, by definition, excluded from the sponsor's financial statement balance sheet, and so it is not reported systematically. Consequently, it is hard to say how extensive the use of SPVs has become. Qualified off-balance sheet SPVs that are used for asset securitization usually issue publicly rated debt and so there is more data about these vehicles. This data is presented and discussed below. SPVs that are not qualified, however, are hidden, as was revealed by the demise of Enron. Enron led to assertions that the use of off-balance sheet SPVs is extreme.[11] But, in fact, the extent of the use of SPVs is unknown.

11. For example, Henry et al. (2002): "Hundreds of respected U.S. companies are ferreting away trillions of dollars in debt in off-balance sheet subsidiaries, partnerships, and assorted obligations."

17.3. SECURITIZATION

Securitization is one of the more visible forms of the use of off-balance sheet SPVs because securitization uses qualified SPVs and involves selling registered, rated securities in the capital markets. Consequently, there is data available. Our empirical work will concentrate on credit card receivables securitization. In this section we briefly review the important features of securitization SPVs.

17.3.1. Overview of Securitization

Securitization involves the following steps: (i) a sponsor or originator of receivables sets up the bankruptcy remote SPV, pools the receivables, and transfers them to the SPV as a "true sale"; (ii) the cash flows are tranched into asset-backed securities, the most senior of which are rated and issued in the market; the proceeds are used to purchase the receivables from the sponsor; (iii) the pool revolves in that over a period of time the principal received on the underlying receivables is used to purchase new receivables; (iv) there is a final amortization period, during which all payments received from the receivables are used to pay down tranche principal amounts. Credit card receivables are different from other pools of underlying loans because the underlying loan to the consumer is a revolving credit; it has no natural maturity, unlike an automobile loan, for example. Consequently, the maturity of the SPV debt is determined arbitrarily by stating that receivable payments after a certain date are "principal" payments.

Figure 17.2 shows a schematic drawing of a typical securitization transaction. The diagram shows the two key steps in the securitization process: pooling and tranching. Pooling and tranching correspond to different types of risk. Pooling minimizes the potential adverse selection problem associated with the selection of the assets to be sold to the SPV. Conditional on selection of the assets, tranching divides the risk of loss due to default based on seniority. Since tranching is based on seniority, the risk of loss due to default of the underlying assets is stratified, with the residual risks borne by the sponsor.

Securitization is a significant and growing phenomenon. Figure 17.3 and Table 17.1 provide some information on non-mortgage QSPV outstanding amounts. The figure shows that the liabilities of non-mortgage vehicles grew rapidly since the late 1990s, and by 2004 amounted to almost $1.8 trillion. Table 17.1 shows the breakdown by type of receivable. Note that credit card receivables are the largest component of (non-mortgage) asset-backed securities. See Kendall and Fishman (1996) and Johnson (2002) for earlier discussions of securitization in the US, and Moody's (May 29, 2003) on the growth of securitization internationally.

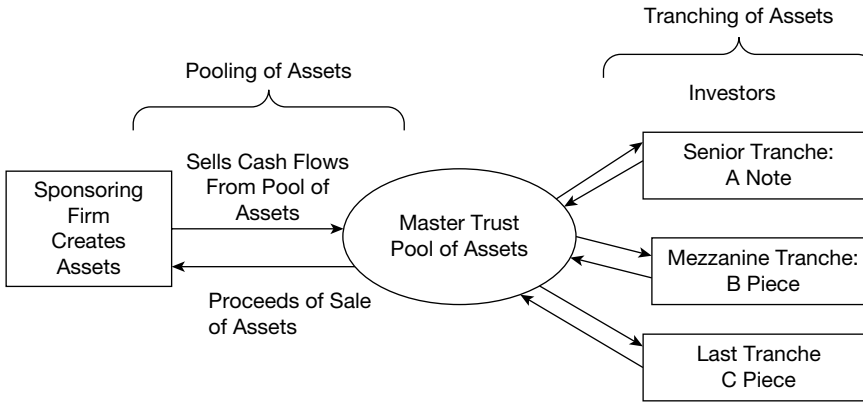


Figure 17.2 Schematic of a Securitization Transaction

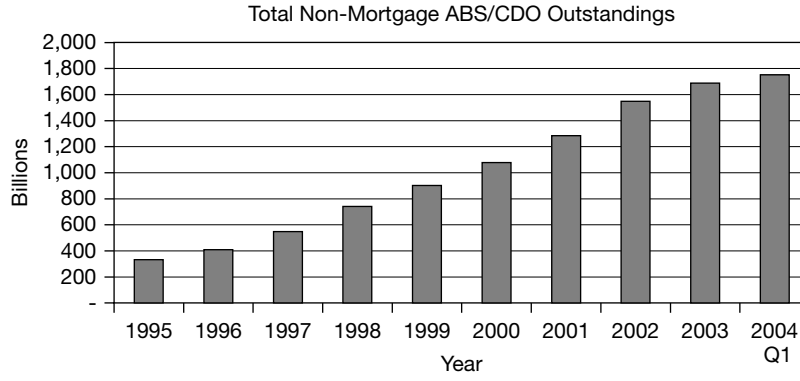


Figure 17.3 Total Non-Mortgage ABS/CDO Outstandings
SOURCE: Bond Market Association

Closely related to securitization is asset-backed commercial paper (ABCP). Asset-backed commercial paper SPVs are called "conduits." ABCP conduits are bankruptcy-remote SPVs that finance the purchase of receivables primarily through issuing commercial paper. ABCP conduits are also very large. The U.S. commercial paper market, as of August 2004, stood at $1.3 trillion, having grown from $570 billion in January 1991. Figure 17.4 shows the ratio of ABCP to total outstanding commercial paper over the last twelve years. Over half of the total consists of ABCP.[12]

12. ABCP conduits are an interesting topic in the own right. See Moody's (1993), FitchIBCA (2001), Elmer (1999), Croke (2003), and Standard and Poor's (2002). ABCP conduits can be multi-seller, meaning that the receivables in the conduit have been originated by different institutions.

Table 17-1. ASSET-BACKED SECURITIES OUTSTANDING AMOUNTS

	Cars	Credit Cards	Home Equity	Manufactured Housing	Student Loans	Equipment Leases	CBO/CDO	Other
1995	59.5	153.1	33.1	11.2	3.7	10.6	1.2	43.9
1996	71.4	180.7	51.6	14.6	10.1	23.7	1.4	50.9
1997	77	214.5	90.2	19.1	18.3	35.2	19	62.5
1998	86.9	236.7	124.2	25	25	41.1	47.6	144.7
1999	114.1	257.9	141.9	33.8	36.4	51.4	84.6	180.7
2000	133.1	306.3	151.5	36.9	41.1	58.8	124.5	219.6
2001	187.9	361.9	185.1	42.7	60.2	70.2	167.1	206.1
2002	221.7	397.9	286.5	44.5	74.4	68.3	234.5	215.4
2003	234.5	401.9	346	44.3	99.2	70.1	250.9	246.8
2004 Q1	238.2	406.5	385.1	43.9	102.4	68.7	253.3	250.4

SOURCE: Bond Market Association.

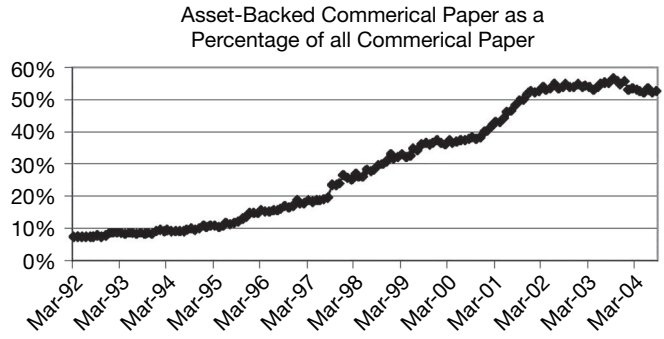


Figure 17.4 Asset-Backed Commercial Paper Conduits
SOURCE: Board of Governors of the Federal Reserve System (http://www.federalreserve.gov/releases/cp/histouts.txt)

17.3.2. The Structure of Securitization Vehicles

Some of the details of the structure of credit-card securitization SPVs are important for the subsequent empirical work. These details are briefly reviewed in this section.

TRUSTS—MASTER TRUSTS

Securitization SPVs are invariably trusts. The sponsor transfers receivables to the trust for the benefit of the certificate holders, i.e., the investors in the SPV. Most trusts are Master Trusts, which allow for repeated transfers of new receivables, whenever the sponsor chooses.[13] At each such instance, the trust issues a series

13. A "discrete trust" is an SPV used for a single initial transfer of assets.

of securities (trust certificates) to investors in the capital markets. Each series has an undivided interest in the assets and an allocable interest in the collections of the receivables in the master trust, based on the size of each series. Trust assets that have not been allocated to a series are called the "seller's interest," discussed below. See Schwarcz (2003a).

Master trusts can be "socialized" or "nonsocialized," two categories which generally refer to how the SPV waterfall works, i.e., how the receivables' cash flows are internally allocated. In nonsocialized trusts there is no reallocation of excess cash flow until each series is paid its full amount. Socialized trusts pay the trust's expenses, including the monthly interest to investors, based on the needs of individualized series. Generally, the socialized excess spread is socialized across all SPV notes issued by the trust. This means that should there be an early amortization event (discussed below), then all the notes go into early amortization. In a nonsocialized trust, the notes have their own separate excess spreads. See Standard and Poor's (n.d.) for details.

Seller's Interest

The "seller's interest" refers to the sponsor's ownership of trust assets that have not been allocated to any series of securities issued by the trust. The size of the seller's interest varies through time as the amounts of securities issued by the SPV changes and as the balance of principal receivables in the trust assets changes. The seller's interest is usually initially set at seven percent.

Excess Spread and Early Amortization

A general feature of asset-backed securities is that they involve "excess spread." The yield on the underlying loans that is paid into the trust should be high enough to cover the payment of interest on the asset-backed securities (ABS) tranches in addition to the servicing fees. Excess spread is generally defined as finance charges collections (i.e., the gross yield on the underlying receivables) minus certificate interest (paid to the holders of the SPV debt), servicing fees (paid to the servicer of the receivables, usually the sponsor), and charge-offs (due to default by the underlying borrowers) allocated to the series. For example:

Gross Yield on Portfolio	18%
Investors' Weighted Avg. Coupon	–7%
Servicing Expense	–2%
Charge-Offs	–5%
Excess Spread	4%

Depending on the structure of the SPV, available excess spread may be shared with other series in the Master Trust, used to pay credit enhancers, deposited into a reserve account to be used to cover charge-offs, or released to the sponsor.

Practitioners view the excess spread as providing a rough indication of the financial health of a transaction. Excess spread is in fact highly persistent and consequently can be used as a way to monitor a transaction.

All credit card structures have a series of early amortization triggers, which if hit cause the payments to investors to be defined as principal, so that the SPVs' liabilities are paid off early, that is, before the scheduled payment date. Early amortization events include insolvency of the originator of the receivables, breaches of representations or warranties, a service default, failure to add receivables as required, and others. Most importantly, however, a transaction will amortize early if the monthly excess spread falls to zero or below for three consecutive months.

Credit Enhancement

In the most common securitization structure the SPV issues tranches of securities to the capital markets based on seniority. There are senior notes, called A notes, and junior or mezzanine notes, called B notes. A common form of credit enhancement to the more senior classes, A notes and B notes, is a subordinated interest known as the collateral invested amount (CIA). The most subordinated interest is referred to by a number of different names, including the C class, C note, or collateral interest.[14] As mentioned above, C notes are typically privately placed. This is partly because they are riskier, but also because they do not qualify as debt for tax purposes making them ERISA-ineligible. Because they are privately placed, they are not rated, and much less information is available about them. See Moody's (November 11, 1994) on C notes.

Credit enhancement for the CIA is a reserve account, which grows depending on the level of the excess spread. If the excess spread is low, then excess spread is trapped inside the SPV and used to build up the reserve account to a specified level. Reserve account structures vary, with different structures having different amounts of excess spread trapped inside the trust depending on different contingencies. If the excess spread is negative, the reserve account is drawn down to make up the shortfall.

17.3.3. Implicit Recourse

There are examples of recourse in credit card securitizations that are known publicly. Moody's (January 1997) gives fourteen examples of "notable instances"

14. Prior to the development and widespread use of CIAs, credit card transactions employed letters of credit (LOCs) from highly rated institutions to protect investors against default. CIAs became prevalent as a way to avoid dependency on the LOC issuer's credit quality.

of voluntary support. The earliest example is from May 1989 and the latest is from November 1996. Higgins and Mason (2004) study a sample of 17 implicit recourse events involving ten banks during the period 1987 to 2001.[15] Higgins and Mason document that firms that engage in subsidization of their SPVs "face long delays before returning to market."

17.4. A Theoretical Analysis of SPVs

In this section we analyze a simple model of off-balance sheet financing, a game played between a representative firm (the sponsor of the SPV) and a large number of investors. The goal is to understand the source of value in the use of SPVs.

For concreteness we call the sponsoring firm a bank, by which we mean any financial intermediary or, indeed, any firm. We proceed by first setting out a model of the bank financing a portfolio of two projects in a one period setting. The bank's efforts determine the quality of the projects, unbeknownst to the lenders to the bank. Project quality is implicitly determined by various activities of banks, including information production, screening, and monitoring, but for simplicity it is modeled as an "effort" choice by the bank.[16] This provides a benchmark against which we can determine the value of securitization in the one period setting.

We will subsequently allow for the possibility of securitization, where one project may be financed off-balance sheet in an SPV. The timing is as follows: projects are allocated to be financed on- or off-balance sheet, and then the bank makes a single effort choice that determines the quality of both the on- and off-balance sheet projects (though ex post their realized qualities can differ). To emphasize, projects are allocated first, and then project quality is realized. So, the focus at this point is on the moral hazard problem involving effort choice, rather than on the strategic allocation of projects after their qualities are known (i.e., the adverse selection problem). By comparing the value of the bank when securitization is allowed to the benchmark bank value when there is no securitization, we determine the factors causing securitization to be valuable.

Finally, we will allow for strategic allocation of the two projects, i.e., projects are allocated between the balance sheet of the bank and the balance sheet of

15. During the period 1987–2001 Higgins and Mason (2004) report two instances of early amortization, both associated with the failure of the sponsoring institution, namely, Republic Bank and Southeast Bank.

16. See Gorton and Winton (2003) for a review of the literature on banks' information production, screening, and monitoring activities.

the SPV *after* their qualities are known. The possibility of strategic allocation of projects adds an additional problem that investors must be concerned about. In this setting, the bank cannot commit to allocate a high type project to the SPV. In the credit card case there are some constraints on the lemons problem because accounts to be sold to the trust are supposed to be chosen randomly. In this case, the adverse selection may have more to do with the timing of the addition of accounts, depending on the state of the on-balance sheet assets, or perhaps with the removal of accounts.[17]

Without the ability to commit to transfer a high quality project to the SPV, we show that no lender will lend to the SPV. Off-balance sheet financing, or securitization, in this setting is not possible. This sets the stage for the repeated SPV game, analyzed briefly in the final part of this section. The point there is that repetition of the stage game between the bank and the outside investors can create equilibria in which an implicit contractual arrangement involving bailouts of the SPV by the sponsoring bank can be enforced. By "bailouts" we mean extra-contractual support for the SPV, as will become clear below.

17.4.1. Model Set-Up

A competitive bank seeks to finance two one-period nondivisible projects. Each project requires \$1 of investment. The bank has an amount $\$E < 2$ available to finance the two projects. Since $E < 2$, the bank must borrow $D = 2 - E$, promising to repay F at the end of the period. Debt, however, is tax advantaged, so only $(1 - \tau)F$ needs to be repaid, where τ is the relevant tax rate. The interest rate in the economy, r, is assumed to be zero for simplicity.

We analyze a representative bank and a unit interval of investors. All agents, i.e., the banks and the investors, are risk-neutral. Consumption occurs at the end of the period.

The bank determines the quality of its projects by expending "effort," $e \in \{e_H, e_L\}$, where $e_H > e_L$, and such that a project returns y^H with probability e and y^L with probability $(1 - e)$, where $y^H > y^L$. The single effort choice determines the qualities of both projects, but project realizations are independent. Thus, there are four possible outcomes or states of the world at the end of the period: $\{y^H, y^H\}$, $\{y^H, y^L\}$, $\{y^L, y^H\}$, and $\{y^L, y^L\}$. The single effort costs $h(e)$. "Effort" is to be interpreted as the resources necessary to produce information about a project and to monitor it. Effort is not contractible.

17. Also, sometimes sponsors add "high quality" accounts to improve the overall quality of the receivables pool.

Projects satisfy the following assumptions:

A1. $2\left[e_H y^H + (1 - e_H)\, y^L\right] - h(e_H) > D$, i.e., a project is a positive net present value investment when a high effort level is chosen, i.e., $e = e_H$.
A2. $2\left[e_L y^H + (1 - e_L)\, y^L\right] - h(e_L) < D$, i.e., a project is a negative net present value investment when a low effort level is chosen, i.e., $e = e_L$.
A3. $2y^L - h(e) < F$, for $e \in \{e_H, e_L\}$, i.e., default is certain if each project returns y^L $\left(\text{state } \left\{y^L, y^L\right\}\right)$.
A4. $2y^H - h(e) > y^H + y^L - h(e) > F$, for $e \in \{e_H, e_L\}$, i.e., default does not occur in the other states.

Assumption A1 ensures that investors will only invest if they are sure that the bank will make a high effort choice. A project is not worth undertaking otherwise. Below, the incentive compatibility constraints ensure that banks will make the high effort choice. Assumptions A3 and A4 are stated in terms of the face value of the debt, F, which is an endogenous variable. Nevertheless, the point of A3 and A4 is to determine the states of the world when default occurs. Default occurs only in the state $\left\{y^L, y^L\right\}$. We will subsequently solve for the equilibrium F under this assumption and then verify that this value of F is consistent with assumptions A3 and A4 when F is eliminated through substitution; the assumptions can then be stated entirely in terms of primitives.

Corporations face a proportional bankruptcy cost, proportional to the realized output. In other words, larger firms have higher bankruptcy costs. This cost is borne by the creditors. Making the bankruptcy cost proportional, rather than lump-sum, is both realistic and also simplifies the model, as will become clear below. The bankruptcy cost is $c \in (0, 1)$ per unit of output. A fixed bankruptcy cost could be added to this, though with binomial outcomes it has no additional content. The bankruptcy cost is discussed further below.

On-balance sheet debt has a tax advantage. Off-balance sheet debt usually does not have this advantage. Here the cost of using off-balance sheet debt is the loss of the tax shield to the sponsoring firm. The sponsor may structure the SPV so that this cost does not exist. In that case, we would point to other costs. In general, some limit to how much can be financed off-balance sheet is needed for there to be an interior solution. However, recent "whole-firm" securitizations suggest that there may be few limits. See Pfister (2000).

17.4.2. Discussion of the Model

The model provides a role for the bank; it has the unique ability to find high quality projects by making an effort. However, this value production is not

observable to outside investors since they cannot confirm the effort level chosen by the bank. This is essentially the usual model of bank activity. We assume that the bank issues debt to outside investors, and do not explain why debt is the security of choice. Any firm transferring assets off-balance sheet has created assets of a certain value, which may not be known to outside investors, so the "bank" need not literally be interpreted to exclude nonfinancial firms.

17.4.3. The Benchmark Case of No Securitization

We begin with the benchmark problem of the bank when there is no off-balance securitization. In that case, the bank's problem is to choose F and $e \in \{e_H, e_L\}$ to maximize the expected value of its projects:

$$\max : V = e^2 \left[2y^H - h(e) - (1-\tau)F\right] + 2e(1-e)\left[y^H + y^L - h(e) - (1-\tau)F\right] \qquad \text{Problem(I)}$$

$$\text{subject to: (i) } E(F) \geq D \qquad \text{(Participation of Investors)}$$

$$\text{(ii) } V(e = e_H; e_0 = e_H) \geq V(e = e_L; e_0 = e_H) \qquad \text{(Incentive Compatibility)}$$

The first constraint says that the expected pay-off to the investors who purchase the bank debt, $E(F)$, must be at least what was lent (D), otherwise the risk neutral investors will not lend to the bank (since the interest rate is zero). The second constraint says that if investors lend to the bank believing that the bank will choose effort level e_H, where e_0 is the belief of the lenders regarding the bank's effort choice, then the bank behaves consistently with these beliefs, choosing $e = e_H$.

The optimization problem is written assuming that the bank defaults only in state $\{y^L, y^L\}$ as assumed above by A3 and A4.

Note that the Participation Constraint can be written as follows, since investors get only the remaining cash flows net of the bankruptcy and effort costs:

$$\left[e^2 + 2e\ (1-e)\right]F + (1-e)^2\left[2y^L(1-c) - h(e)\right] \geq D.$$

Suppose investors' beliefs about the bank's effort choice are $e = e_0$. Then the lowest promised repayment amount that lenders will accept, in order to lend, is:

$$F_0 = \frac{D - (1-e_0)^2\left[2y^L(1-c) - h(e_0)\right]}{e_0\ (2-e_0)}.$$

Substituting this into the bank's problem, the bank's problem is now to choose $e \in \{e_H, e_L\}$ to:

$$\max V = 2ey^H + 2e(1-e)y^L - e(2-e)h(e) - (1-\tau)e(2-e)\left[\frac{D-(1-e_0)^2\left[2y^L(1-c)-h(e_0)\right]}{e_0(2-e_0)}\right]$$

$$\text{subject to :} \quad \text{(ii)}\ V(e=e_H;\ e_0=e_H) \geq V(e=e_L;\ e_0=e_H) \quad \text{(Incentive Compatibility).}$$

Incentive compatibility requires that the bank's choice of $e \in \{e_H, e_L\}$ be the same as what the lenders believe it will be, namely e_0. Suppose that beliefs are consistent, i.e., that $e = e_0 = e_H$. Then, indicating bank value by V^H, we have:

$$V^H = 2e_H y^H + 2e_H(1-e_H)y^L - e_H(2-e_H)h(e_H) - (1-\tau)\left[D-(1-e_H)^2\left(2y^L(1-c)-h(e_H)\right)\right] \qquad (17.1)$$

If beliefs were inconsistent, that is, if lenders' beliefs were $e_0 = e_H$ but the bank chose $e = e_L$, then the value of the bank would be given by:

$$V(e=e_L;\ e_0=e_H) = 2e_L y^H + 2e_L(1-e_L)y^L - e_L(2-e_L)h(e_L) - (1-\tau)e_L(2-e_L)\left[\frac{D-(1-e_H)^2\left(2y^L(1-c)-h(e_H)\right)}{e_H(2-e_H)}\right]$$

LEMMA 1: If:

$$2y^H(e_H-e_L) + 2y^L\left[e_H(1-e_H)-e_L(1-e_L)\right] - h(e_H)e_H(2-e_H) + h(e_L)e_L(2-e_L) - (1-\tau)\left[D-(1-e_H)^2\right.$$

$$\left.\left[2y^L(1-c)-h(e_H)\right]\left[1-\frac{e_L(2-e_L)}{e_H(2-e_H)}\right] > 0,\right.$$

then at the optimum, investors believe $e_0 = e_H$ and the bank chooses $e = e_H$. The value of the bank is given by (17.1).

Proof: The incentive compatibility constraint, $V(e=e_H;\ e_0=e_H) \geq V(e=e_L;\ e_0=e_H)$, is satisfied if the condition in the lemma holds. It remains to verify that the equilibrium F derived under A3 and A4 is consistent, i.e., to state A3 and A4 in terms of primitives. That is left to the Appendix. //

In what follows we will refer to V^H as the value of the bank when there is no securitization. This will be the benchmark value against which the value of the bank with securitization will be compared.

17.4.4. Special Purpose Vehicles and Securitization

Now, suppose the bank sets up a special purpose vehicle (SPV) to finance one of the projects. One project will be financed on-balance sheet, and one will be financed off-balance sheet.[18] The SPV has no bankruptcy costs, as discussed above, and its debt has no tax advantage. As before, the effort choice is made at the bank level and determines the qualities of both projects, though the outcomes are independent.[19] To be clear, the projects are first allocated to be on- or off-balance sheet, and then the bank makes its effort choice.

On-balance sheet the bank will borrow 0.5D, promising to repay F^B at the end of the period. Off-balance sheet, the SPV will borrow 0.5D, promising to repay F^S at the end of the period.[20] The bank then has two assets on-balance sheet, its own project, and an equity claim on the SPV, i.e., if y is the realization of the SPV's project, then the bank's equity claim on the SPV at the end of the period is $\max\left[y - F^S, 0\right]$.[21]

Assumptions analogous to A3 and A4, above, define the bankruptcy states:

A3a. $2y^L - h(e) < F^B + F^S$, for $e \in \{e_H, e_L\}$, i.e., default of both the bank and the SPV occurs if the realized state of the world is $\left\{y^L, y^L\right\}$.

A4a. $2y^H - h(e) > y^H + y^L - h(e) > F^B + F^S$, for $e \in \{e_H, e_L\}$, i.e., there need not be default of either entity in the other states.

As before, assumptions A3a and A4a are stated in terms of F^B and F^S, endogenous variables. Assumption A3a determines the states of the world when default definitely will occur, namely, in state $\left\{y^L, y^L\right\}$. A4a states that the two projects generate sufficient payoffs in the other states to avoid bankruptcy, though whether that is the outcome or not will depend on the relationship between the bank and the SPV. We will subsequently solve for the equilibrium F^B and F^S under these assumptions and then verify that those values of F^B and F^S are consistent with assumptions A3a and A4a when F is eliminated through substitution; the assumptions can then be stated entirely in terms of primitives.

18. This assumption is made for simplicity. The model does not determine the scale of the SPV.

19. Note that no effort choice can be made by the SPV, as it is passive. If the effort choice could be made at that level, the entity would be a subsidiary of the bank, rather than an SPV.

20. For simplicity other financing choices are assumed to not be available. While we do not model tranching, it is not inconsistent with the model to allow for additional motivations for securitization beyond those we consider, such as clientele effects (e.g., perhaps due to ERISA-eligibility requirements).

21. Strictly speaking there is an intermediate step because the bank funds both projects initially on-balance sheet and then transfers one, in a true sale, to the SPV. We assume that the proceeds from selling the project to the SPV are used to pay down on-balance sheet debt. For simplicity, this step is omitted.

We also now assume:

A5. $(1-e_H)^2 y^L (1-c) < 0.5D$, i.e., the expected return for the bank, from the on-balance sheet project in the bankruptcy state $\{y^L, y^L\}$ (which occurs with probability $(1-e_H)^2$), is insufficient to pay 0.5D, the amount borrowed.

At the end of the period, by A3a and A4a, the possible outcomes are as follows, where the first element is the on-balance sheet project state realization and the second element is the off-balance sheet project state realization:

- $\{y^H, y^H\}$: Both projects realize y^H; this occurs with probability e^2, $e \in \{e_H, e_L\}$. In this event, both on- and off-balance sheet debts can be repaid in full.
- $\{y^H, y^L\}$: The off-balance sheet project realizes y^H, but the SPV's project is worth y^L. This occurs with probability $e(1-e)$, $e \in \{e_H, e_L\}$. The bank is solvent, but the SPV defaults on its debt.
- $\{y^L, y^H\}$: The off-balance sheet project realizes y^H, but the bank's project is worth y^L. This occurs with probability $e(1-e)$, $e \in \{e_H, e_L\}$. The SPV can honor its debt, and so can the bank because the bank is the equity holder of the SPV.
- $\{y^L, y^L\}$: Both projects realize y^L; this occurs with probability $e(1-e)$, $e \in \{e_H, e_L\}$. Neither the bank nor the SPV can honor their debt.

Note that with or without securitization, the bank fails only if the realized state is $\{y^L, y^L\}$. Consequently, with only two states a lump-sum bankruptcy cost would always be borne in this, and only this, state. This is due to the simplicity of the model. However, the proportional bankruptcy cost will be affected by securitization since the on-balance sheet assets have been reduced to one project. In a more complicated model, with a continuous range of project realizations, a fixed bankruptcy cost could be borne as a function of the bank's leverage, which could be chosen endogenously. Here, the simplicity of the model dictates use of a proportional bankruptcy cost. But, clearly this is not essential for the main point.

The bank's problem is to choose F^B, F^S, and $e \in \{e_H, e_L\}$ to:

$$\begin{aligned}\max V^S = {} & e^2\left[2y^H - h(e) - (1-\tau)F^B - F^S\right] \\ & + e(1-e)\left[y^L + y^H - h(e) - (1-\tau)F^B - F^S\right] \\ & + e(1-e)\left[y^H - h(e) - (1-\tau)F^B\right] \quad \text{Problem (II)}\end{aligned}$$

$$\text{s.t. (i)} \quad E\left[F^B\right] \geq 0.5D \qquad \left(\text{Participation of Investors in the Bank}\right)$$
$$\text{(ii)} \quad E\left[F^S\right] \geq 0.5D \qquad \left(\text{Participation of Investors in the SPV}\right)$$
$$\text{(iii)} \quad V^S\left(e = e_H;\ e_0 = e_H\right) \geq V^S\left(e = e_L;\ e_0 = e_H\right) \left(\text{Incentive Compatibility}\right)$$

The solution method for Problem (II) is analogous to that for Problem (I), and so is left to the Appendix (including a lemma, Lemma 2, that is analogous to Lemma 1.) We refer to V^S as the resulting value of the bank with securitization. We now state:

PROPOSITION 1 (FEASIBILITY OF SECURITIZATION). If $(1 - e_H)^2 y^L c - \tau[0.5D - (1 - e_H)^2 y^L (1 - c)] > 0$, then it is optimal for the bank to use the SPV to finance one project.

Proof: The condition in the proposition is a simplification of $V^S - V^H > 0$. //

The factors that effect the profitability of securitization are taxes (τ), the bankruptcy cost (c), and risk, as measured by $(1 - e_H)^2$, i.e., the chance of bankruptcy occurring. Taxes matter, to the extent that bankruptcy does not occur, because debt issued by the SPV is not tax advantaged (by assumption). The bankruptcy cost matters because expected bankruptcy costs are reduced to the extent that projects are financed off-balance sheet. This is due to the legal structure of the SPV. Finally, the risk of bankruptcy, $(1 - e_H)^2$, makes the chance of incurring the bankruptcy cost higher.

COROLLARY 1: The profitability of off-balance sheet financing is increasing in the bankruptcy cost, c, decreasing in the tax rate, τ, and increasing in the riskiness of the project (i.e., the chance of bankruptcy),$(1 - e_H)^2$.

Proof: The derivatives of $V^S - V^H$ with respect to c, τ, and $(1 - e_H)^2$, respectively, are:

$$\frac{\partial(V^S - V^H)}{\partial \tau} = -\left[0.5D - (1 - e_H)^2 y^L (1 - c)\right] < 0, \text{ by A5.}$$
$$\frac{\partial(V^S - V^H)}{\partial c} = (1 - e_H)^2 y^L (1 - \tau) > 0.$$
$$\frac{\partial(V^S - V^H)}{\partial (1 - e_H)^2} = (1 - \tau) c y^L + \tau y^L > 0. \ //$$

Corollary 1 identifies the basic drivers of SPV value, under the assumption that the projects are allocated to on- or off-balance sheet before their quality if known, i.e., there is no adverse selection.

17.4.5. Securitization with Moral Hazard and Strategic Adverse Selection

Now, suppose that the bank makes an effort choice, i.e., $e \in \{e_H, e_L\}$, but then *after* observing the realized project qualities, one of the projects is allocated to the SPV. Recall that project quality is not verifiable. This means that investors in the debt issued by the SPV face an additional problem. In addition to the moral hazard associated with the effort choice, there is an adverse selection problem with regard to which project is allocated to the SPV, the strategic adverse selection problem.

For this subsection we will also assume:
A6. $e_H{}^2 y^H + \left(1 - e_H{}^2\right) y^L < 0.5D$.

The meaning of A6 will become clear shortly.

With the possibility of strategic adverse selection, at the end of the period the possible outcomes (following A3a and A4a) are as follows:

- $\left\{y^H, y^H\right\}$: Both projects realize y^H; this occurs with probability e^2. The bank allocates one of the y^H projects to the SPV and retains the other one on-balance sheet. Both on- and off-balance sheet debts can be repaid in full.
- $\left\{y^H, y^L\right\}$ and $\left\{y^L, y^H\right\}$: The realization of projects is: one y^H and one y^L. This occurs with probability $2e(1–e)$. In both of these states of the world, the bank keeps the y^H project on-balance sheet and allocates the y^L project to the SPV. The bank is solvent, but the SPV defaults on its debt.
- $\left\{y^L, y^L\right\}$: Both projects realize y^L; this occurs with probability $(1-e)^2$. One of the y^L projects is allocated to the SPV and the bank retains the other on-balance sheet. Neither the bank nor the SPV can honor its debt.

In the previous subsection the SPV failed in two states of the world, the two situations where it realized y^L. Now, the SPV fails in three states of the world, due to the strategic adverse selection problem. Only if $\left\{y^H, y^H\right\}$ is realized will the SPV be solvent. So, the expected income of the SPV is: $e^2 y^H + \left[2e(1-e) + (1-e)^2\right] y^L = e^2 y^H + \left(1 - e^2\right) y^L$. But this is less than 0.5D, by A6. Consequently, no investor will lend to the SPV. Recognizing this problem, the bank would like to commit to not engage in strategic adverse selection; the bank would like to commit to allocate projects prior to the realization of the project outcome. But there is no way to do this because project quality is not verifiable.

Imagine for a moment that the bank could commit to subsidize the SPV in the event that the SPV realized y^L and the bank realized y^H. Shortly, we will make

clear what "subsidize" means. Let F^{SC} be the face value of the debt issued by the SPV under such commitment, and F^{C} the corresponding face value of the debt issued by the bank. Then at the end of the period, the possible outcomes would be as follows:

- $\{y^H, y^H\}$: Both projects realize y^H; this occurs with probability e^2. Both on- and off-balance sheet debts can be repaid in full. The expected profit to the bank in this case is:

$$e^2 \left[2y^H - h(e) - (1-\tau) F^C - F^{SC}\right].$$

- $\{y^H, y^L\}$: The bank's project is worth y^H and the SPV's is worth y^L. This occurs with probability $e(1-e)$. The bank is solvent and subsidizes the SPV, so that neither defaults on its debt. "Subsidize" means that the bank assumes responsibility for the debt of the SPV. The bank's expected profit in this state of the world is:

$$e(1-e)[y^H + y^L - h(e) - (1-\tau)F^C - F^{SC}].$$

- $\{y^L, y^H\}$: The bank's project is worth y^L and the SPV's is worth y^H. This occurs with probability $e(1-e)$. The SPV is solvent. Without the return on its SPV equity the bank would be insolvent. But the SPV has done well so that neither defaults on its debt. The expected profit in this case is the same as in the previous case, though the interpretation is different:

$$e(1-e)[y^H + y^L - h(e) - (1-\tau)F^C - F^{SC}].$$

- $\{y^L, y^L\}$: Both projects realize y^L; this occurs with probability $(1-e)^2$. Neither the bank nor the SPV can honor its debt. The bank earns zero.

With this commitment, the bank's problem is to choose F^C, F^{SC}, and $e \in \{e_H, e_L\}$ to:

$$\max V^C = e^2[2y^H - h(e) - (1-\tau)F^C - F^{SC}] + 2e(1-e)$$
$$[y^H + y^L - h(e) - (1-\tau)F^C - F^{SC}] \quad \text{Problem (III)}$$

s.t. (i) $E[F^C] \geq 0.5D$ (Participation of Bank Investors)

(ii) $E[F^{SC}] \geq 0.5D$ (Participation of SPV Investors)

(iii) $V^C(e = e_H;\ e_0 = e_H) \geq V^C(e = e_L;\ e_0 = e_H)$ (Incentive Compatibility)

Constraints (i) and (ii) can be re-written, respectively, as:

$$e(2-e)F^{C} + (1-e)^{2}[y^{L}(1-c) - h(e)] \geq 0.5D,$$

and

$$e(2-e)F^{SC} + (1-e)^{2}y^{L} \geq 0.5D.$$

The solution to Problem (III) is contained in the Appendix, including a lemma, Lemma 3, that is analogous to Lemma 1. We refer to V^{C} as the resulting value of the bank with commitment. We now state:

PROPOSITION 2 (EQUIVALENCE OF PROBLEMS II AND III). If the bank can commit to subsidize the SPV, then the profitability of the bank is the same as it would be when projects were allocated between the bank and the SPV prior to their realizations, i.e., when there was no strategic adverse selection.

Proof: It may be verified that $V^{S} = V^{C}$. //
Intuitively, while the debt is repriced to reflect the subsidy from the bank in the state $\{y^{H}, y^{L}\}$, there are no effects involving the bankruptcy cost or taxes. Consequently, the bank's value is the same as in problem II when projects were allocated between the bank and the SPV prior to their realizations.

Proposition 2 states that securitization would be feasible, i.e., investors would lend to the SPV, and it would be profitable for the bank (under the conditions stated in Proposition 1), if it were possible to overcome the problem of strategic adverse selection by the bank committing to subsidize the SPV. However, accounting and regulatory rules prohibit such a commitment, even if it were possible. That is, a formal contract, which can be upheld in court and which is consistent with accounting and regulatory rules, effectively would not be consistent with the SPV being a QSPV, and hence the debt would not be off-balance sheet. The bankruptcy costs would not be minimized. We now turn to the issue of whether a commitment is implicitly possible in a repeated context.

17.4.6. The Repeated SPV Game: The Implicit Recourse Equilibrium

In any single period, the bank cannot securitize a project because lenders will not lend to the SPV due to the strategic adverse selection problem. We now consider an infinite repetition of the one period problem, where for simplicity we assume that the bank has exactly $E available every period to finance the two projects.[22] The one-shot-game outcome of no securitization can be infinitely repeated, so

22. In other words, we assume that if the bank does well it pays a dividend such that E remains as the equity in the bank. If the bank does poorly, we assume that the bank can obtain more equity so that again there is E. Obviously, this omits some interesting dynamics about the bank's capital

this is an equilibrium of the repeated game. However, the idea that repetition can expand the set of equilibria, when commitment is possible, is familiar from the work of Friedman (1971), Green and Porter (1984), and Rotemberg and Saloner (1986), among others. The usual context is oligopolistic competition, where the competing firms are incompletely informed about their rivals' decisions. The firms want to collude to maintain oligopolistic profits, but cannot formally commit to do so. Here the context is somewhat different. The sponsoring bank and the investors in the SPV "collude" in adopting a contractual mechanism that cannot be written down because of accounting and regulatory rules. In a sense the two parties are colluding against the accountants and regulators. We will call such an equilibrium an "Implicit Recourse Equilibrium."

For this section we will suppose that the interest rate, r, is positive and constant. This means that everywhere there was a "D" above, it must be replaced by $(1 + r)\,D$, as the risk neutral investors require that they earn an expected rate of return of r.

The basic idea of repeating the SPV game is as follows. Suppose investors believe that the bank will subsidize the SPV in the state $\{y^H, y^L\}$, when the SPV would otherwise default. That is, investors have priced the debt as F^C and F^{SC}, as given above, and their beliefs are $e_0 = e_H$. Now, suppose that the state $\{y^H, y^L\}$ occurs, that is, the state of the world where the bank is supposed to subsidize the SPV. The realized bank profit is supposed to be:

$$y^H + y^L - h(e_H) - (1 - \tau)\,F^C - F^{SC}.$$

But, suppose the bank reneges and leaves the SPV bankrupt with $y^L - F^{SC} < 0$, i.e., there is no subsidy. The SPV then defaults on its debt. In that case, on-balance sheet the bank realizes:

$$y^H - h\,(e_H) - (1 - \tau)F^C.$$

So, the one-shot gain from reneging on the implicit contract is $F^{SC} - y^L > 0$. Since this is positive, the bank has an incentive to renege. But, in a repeated setting, investors can punish the bank by not investing in the bank's SPV in the future, say for N periods. If the bank cannot securitize again for N periods, it loses (from Proposition 1):

$$\sum_{t=1}^{N} \delta^t (V^S - V^H) = \sum_{t=1}^{N} \delta^t [(1 - e_H)^2 y^L c - \tau c (1 - e_H)^2 y^L - \tau [0.5D - (1 - e_H)^2 y^L]],$$

ratio and begs the question of the coexistence of outside equity and debt. These issues are beyond the scope of this paper.

where δ is the discount rate. Obviously, the bank will not renege on subsidizing the SPV if the expected present value of the loss is greater than the one-shot gain to deviating. There are combinations of N and δ that will support the Implicit Recourse Equilibrium. While this is the intuition for Implicit Recourse Equilibrium, it clearly depends on the beliefs of the investors and the bank. There may be many such equilibria, with very complicated, history dependent, punishment strategies.

The idea is for the investors in the SPV to enforce support when needed by the threat of refusing to invest in SPV debt in the future if the sponsoring firm deviates from the implicit contract. This means that there is a "punishment period" where investors refuse to invest in SPV debt if the sponsor has not supported the SPV in the past. In general, strategies can be path dependent in complicated ways (See Abreu (1988)). However, a simple approach is to restrict attention to punishments involving playing the no-SPV stage game equilibrium for some period of time, starting the period after a deviation has been detected. We adopt this approach and assume investor and bank beliefs are consistent with this.

For simplicity we will construct a simple example of an Implicit Recourse Equilibrium. Assume that all agents discount at the rate r, and consider the case where $N = \infty$. This corresponds to a "punishment period" of forever.[23] At the start of each period the game proceeds as follows:

1. The bank and the SPV offer debt in the capital markets to investors with face values of F^C and F^{SC}, respectively.
2. Investors choose which type of debt, and how much, to buy.

If investors purchase the SPV debt, then off-balance sheet financing proceeds. Otherwise the bank finances both projects on-balance sheet.

At the end of a period, the state of the world is observed, but cannot be verified. If the state of the world is $\{y^H, y^L\}$, i.e., the on-balance sheet project returns y^H while the off-balance sheet project returns y^L, then the bank is supposed to subsidize the SPV, as described above. At the start of any period, both the banks and investors know all the previous outcomes.

Consider the following trigger strategy based on investor and bank beliefs: If the bank ever does not subsidize the SPV when the state of the world is $\{y^H, y^L\}$, then investors never again invest in the SPV because they believe that the sponsor will not support it and hence the promised interest rate, corresponding to F^{SC}, is too low. The bank believes that if it deviates investors will never again buy its SPV's debt in the market. Then a subgame perfect Nash equilibrium exists under certain conditions:

23. We do not claim that this is the optimal punishment period.

PROPOSITION 3 (EXISTENCE OF THE IMPLICIT RECOURSE EQUILIBRIUM). If there exists an interest rate, $0 \leq r \leq 1$, such that the following quadratic inequality is satisfied,

$$0.5Dr^2 + r\left\{0.5D\left[1 - \tau e_H\,(2 - e_H)\right] + (1 - e_H)^2 h(e_H) + y^L B\right\} - 0.5D\tau e_H(2 - e_H) + y^L A > 0$$

where $A \equiv [(1 - e_H)^2(c + \tau(1 - c))e_H(2 - e_H) - \tau(1 - e_H)^2 ce_H(2 - e_H)]$
and $B \equiv [(1 - e_H)^2(1 - c) - e_H(2 - e_H)]$,
then securitization is feasible and optimal for any bank that would choose securitization were it able to commit to the policy of subsidization.

Proof: See Appendix.

Obviously, other equilibria could exist. But, the point is that there can exist equilibria where the costs of bankruptcy are avoided by using off-balance sheet financing.

17.4.7. Summary and Empirical Implications

The conclusion of the above analysis is that the value of SPVs lies in their ability to minimize expected bankruptcy costs—securitization arises to avoid bankruptcy costs. By financing the firm in pieces, control rights to the business decisions are separated from the financing decisions. The sponsor maintains control over the business while the financing is done via SPVs that are passive; that is, there are no control rights associated with the SPVs' assets. Bankruptcy is a process of transferring control rights over corporate assets. Off-balance sheet financing reduces the amount of assets that are subject to this expensive and lengthy process.

We have argued that the ability to finance off-balance sheet via the debt of SPVs is critically dependent on a relational, or implicit, contract between the SPV sponsor and investors. The relational contract depends upon repeated use of off-balance sheet financing. We showed that this repetition can lead to an equilibrium with implicit recourse. Such an equilibrium implements the outcome of the equilibrium with formal commitments (Problem III), were such contracts possible. The comparative static properties of the Implicit Recourse Equilibrium are based on the result that the equilibrium outcomes of the Implicit Recourse Equilibrium are the same as the commitment equilibrium.

The idea of a relational contract supporting the feasibility of SPVs leads to our first set of empirical tests, namely, that the trigger strategy can only provide intertemporal incentives for the sponsor insofar as the sponsor exists. If the sponsor is so risky that there is a chance the sponsor will fail, and be unable to support

the SPV, then investors will not purchase the SPV debt. We examine this idea by testing the hypothesis that investors, in pricing the debt of the SPV, care about the risk of the sponsor defaulting, above and beyond the risks of the SPV's assets.

The second hypothesis that we empirically investigate is suggested by Corollary 1. Because the Implicit Recourse Equilibrium implements the outcome with formal commitment, Corollary 1 also describes the repeated equilibrium with implicit recourse. Corollary 1 says that the profitability of off-balance sheet financing is increasing in the bankruptcy cost, c, and increasing in the riskiness of the project (i.e., the chance of bankruptcy),$(1 - e_H)$. In other words, riskier sponsors should securitize more, *ceteris paribus.* Bankruptcy costs are not observable, but the riskiness of the firm can be proxied for by its firm bond rating.

17.5. DATA

The rest of the paper empirically examines these two hypotheses. Our analysis suggests that the risk of a sponsoring firm should, because of implicit recourse, affect the risk of the ABS that are issued by its SPVs. We measure the sponsor's risk by its bond rating, and focus on two ways that this risk might be manifested. As mentioned above, we first consider whether investors care about the strength of the sponsoring firm, above and beyond the characteristics of the ABS themselves. Second, we consider whether riskier firms are more likely to securitize in the first place. To these ends we utilize a number of datasets.

To investigate our first topic, investors' sensitivity to the sponsor's strength, we obtained from Moody's a unique dataset describing every credit-card ABS issued between 1988:06 and 1999:05 that Moody's tracked. This covers essentially all credit-card ABS through mid-1999. The dataset includes a detailed summary of the structure of each ABS, including the size and maturity of each ABS tranche. It summarizes the credit enhancements behind each tranche, such as the existence of any letters of credit, cash collateral accounts, and reserve accounts. Moody's also calculated the amount of direct subordination behind each A and B tranche. [24] These variables contain the information about the ABS structure that investors observed at the time of issuance. Further, the dataset

24. The amount of subordination behind the A note is calculated as (BalB + BalC)/(BalA + BalB + BalC), where BalX is the size (the balance) of tranche X when it exists. The dataset provided the current amount of subordination using current balances. For our analysis below, we want the original amount of subordination at the time of issuance. We were able to estimate this given the original balance sizes of the A and B notes, as well as an estimate of the size of any C note. The size of C notes is not directly publicly available, but we backed out their current size from the reported current amount of subordination behind the B notes. We used this to estimate the original amount of subordination behind the A and B notes.

includes some information about the asset collateral underlying each ABS, such as the age distribution of the credit-card accounts. Also included is the month-by-month ex post performance of each note, in particular the excess spread and its components like the chargeoff rate. The sample used below includes only the A and B tranches, i.e., the tranches that were sold publicly.

Although it is difficult to find pricing information on credit-card ABS, we obtained from Lehman Brothers a dataset containing the initial yields on a large subset of these bonds that were issued in 1997–1999, for both the A and B notes. We obtained similar data from *Asset Sales Reports* for bonds that were issued before 1997. We computed the initial spread as the initial yield minus one month LIBOR at the time of issuance. We also collected Moody's ratings from Bloomberg for the sponsors of each ABS in the Moody's dataset above, which are typically banks. We use the bank's senior unsecured bond rating at issuance.[25]

To investigate our second topic, an analysis of which banks securitize, we use the bank ("entity") -level *Call Report* panel data that comes from the regulatory filings that banks file each quarter, from 1991:09 to 2000:06. Before 1996 we use only the third quarter (September) data, since credit card securitizations were reported only in the third quarter during that period. We also obtained from Moody's a large dataset of all of their ratings of banks' long-term senior obligations, including an ID variable that allowed us to match this data to the *Call Report* ID variables. Accordingly our sample includes all the banks in the *Call Report* dataset for which we have a matching rating.[26] This yields a sample of almost 400 banks and over 5000 bank-quarters, which is large relative to the samples analyzed in previous related literature.

17.6. EMPIRICAL TESTS: ARE THERE IMPLICIT RECOURSE COMMITMENTS?

In this section we analyze the determinants of the spread on the notes issued by the SPVs to the capital markets. Borgman and Flannery (1997) also analyze asset-backed security spreads, over the period 1990–1995. They find that credit card ABS require a lower market spread if the sponsoring firm is a bank or if the sponsor includes guarantees as a form of credit enhancement.

The unit of observation is a transaction, that is a note issuance, either the A note or the B note. We examine the cross sectional determinants of the spreads.

25. We use the rating of the current owner of the ABS trust, accounting for any mergers and acquisitions.

26. Since small banks are less likely to be rated, matches are most common for the larger banks.

The spreads provide us with investors' assessment of the risk factors behind each note. All the A notes were on issuance rated AAA by Moody's.[27] If these ratings are sufficient statistics for default, then the probability of default should be the same for all the A notes and in the simplest case (e.g., if there is no implicit recourse) presumably investors would pay the same initial price for them. Even if there are differences across notes in the quality of the underlying assets or in other factors, the securitizations should be structured to offset these differences and yield the same probability of default. As discussed above, to test for the existence of a relational contract allowing for recourse, we examine whether other factors affect the initial prices of the notes, in particular whether the strength of the sponsor matters, as estimated by its senior unsecured credit rating at the time of issuance. Specifically, we estimate equations of the following form:

$$\begin{aligned}\text{Spread}_{i,j,k,t} = &\beta_0{}'\text{Time}_t + \beta_1{}'\text{Structure}_i + \beta_2{}'\text{Assets}_i \\ &+ \beta_3{}'\text{Trust}_j + \beta_4{}'\text{Rating}_{k,t} + \varepsilon_{i,j,k,t}, \qquad (17.2)\end{aligned}$$

where $\text{Spread}_{i,j,k,t}$ is the initial spread (net of one month LIBOR) on note i from trust j and sponsor k at the time t of issuance. **Time** is a vector of year dummies that control for time varying risk premia as well as all other macroeconomic factors, including the tremendous growth in the ABS market over the sample period. $\textbf{Structure}_i$ represents the structure of tranche i at the time of issuance, such as the degree of subordination and other credit enhancements supporting it, and $\textbf{Assets}_i$ represents the quality of the credit-card assets underlying the tranche at that time. $\textbf{Trust}_j$ is a vector of trust dummies. $\textbf{Rating}_{k,t}$ is the senior unsecured bond rating of the sponsor k of the notes' trust at the time of issuance. The trust dummies control for all trust fixed effects. Since many sponsors have multiple trusts, the dummies also essentially control for sponsor fixed effects.[28] Given this, the ratings variable will essentially capture the effect of changes in a sponsor s rating over time.[29]

Our initial sample includes only the A notes, but later we add the B notes, with **Structure** then including an indicator for the B notes (Junior). Table 17.2 presents summary statistics for the key variables used in the analysis, for the sample of A notes. The sample runs from 1988–1999. Over that time the average A-note spread was just under 50 basis points (b.p.), with a relatively large standard deviation of 68 b.p. About half of the sponsors have ratings of single

27. All but two of the B notes were initially rated A; the two exceptions were rated AA. By distinguishing the A- and B-notes, the analysis implicitly controls for any clientele effects.

28. Though a given trust can also have multiple owners over time, e.g. after a merger or acquisition.

29. As evidenced by the significant results below, there is substantial within-trust variation in both the spreads and ratings over time, with over 30% of trusts exhibiting some change in rating over the sample period.

Table 17-2. SPONSOR RATINGS AND INITIAL SPREADS ON A NOTES: SUMMARY STATISTICS

	Mean	**s.d.**
Spread	0.48	0.68
RatingAA	0.25	0.44
RatingA	0.49	0.50
RatingB	0.26	0.44
LowSub	0.25	0.44
Maturity	5.70	2.25
SellersInt	6.38	1.21
Fixed Rt	0.35	0.48
I_CCA	0.43	0.50
I_LOC	0.03	0.17
I_RES	0.01	0.08
I_Other	0.02	0.15
Seasoned	0.43	0.50
Chargeoff	5.35	1.86

NOTES: N = 167. The sample is that for A Notes in Table 17.3 column (5), averaging over 1988–99.

A (RatingA) on their senior unsecured debt, with the rest being about equally likely to have ratings of AA (RatingAA) or ratings of Baa and Ba (RatingB).

17.6.1. Analysis of the A-Note Spreads

Table 17.3 shows the results for the A notes. Column (1) includes only the year dummies (omitting 1988[30]) and the sponsor ratings (as well as the trust fixed effects). Nonetheless, the adjusted R^2 is already relatively large. The year dummies are significant, with spreads peaking in the early 1990s, perhaps due to the recession. The sponsor ratings at the bottom of the table are of primary interest. Relative to the omitted AA-rated sponsors, the effects of riskier sponsor ratings are positive and monotonic. The coefficient on RatingB for the riskiest (Baa and Ba) sponsors is statistically significant. Thus investors do indeed require higher yields for bonds issued by the trusts of riskier sponsors. That is, even though the A notes all have the same bond ratings, the strength of the sponsor also matters, consistent with our model. This effect is also economically significant. The riskiest sponsors must pay an *additional* 46 b.p. on average, which is about the same size as the average A-note spread and sizable relative to the standard deviation of spreads in Table 17.2. This is a relatively strong result given the trust dummies

30. Because of missing values in some of the covariates, some of the time dummies drop out of the regressions.

Table 17-3. SPONSOR RATINGS AND INITIAL SPREADS ON A NOTES

	(1)		(2)		(3)		(4)		(5)	
	coef.	**t**	**coef.**	**t**	**coef.**	**t**	**coef.**	**t**	**coef.**	**t**
Yr89	−0.565	−0.92	—							
Yr90	—		—		—		—		—	
Yr91	0.915	2.79	1.263	2.13	0.339	0.73	1.360	2.82	0.671	1.34
Yr92	0.886	1.72	-		-		-		-	
Yr93	0.275	0.77	1.456	3.96	—		1.037	3.13	0.491	1.29
Yr94	−0.004	−0.01	0.069	0.24	−0.804	−3.26	0.216	0.85	0.034	0.11
Yr95	−0.771	−2.32	−0.150	−0.56	−1.155	−4.81	−0.137	−0.57	−0.409	−1.44
Yr96	−0.903	−2.78	−0.196	−0.74	−1.091	−4.44	−0.080	−0.34	−0.456	−1.70
Yr97	−0.819	−2.52	−0.132	−0.54	−1.126	−4.77	−0.106	−0.48	−0.519	−2.07
Yr98	−0.940	−2.84	−0.302	−1.33	−1.274	−5.44	−0.262	−1.26	−0.502	−2.27
Yr99	−0.659	−1.60	—		−1.019	−3.52	—		—	
LowSub			0.398	2.81	0.147	1.29	0.136	1.14	0.173	1.57
Maturity					0.050	3.20	0.049	3.10	0.039	2.56
SellersInt					−0.030	−0.39	−0.027	−0.33	0.004	0.06
FixedRt					0.713	8.67	0.722	8.09	0.726	9.05
I_CCA							−0.066	−0.39		

Table 17-3. (CONTINUED)

	(1)		(2)		(3)		(4)		(5)	
	coef.	**t**	**coef.**	**t**	**coef.**	**t**	**coef.**	**t**	**coef.**	**t**
I_LOC							−0.107	−0.28		
I_RES							−0.228	−0.46		
I_Other							0.014	0.06		
Seasoned									−0.331	−2.92
Chargeoff									0.098	2.48
RatingA	0.235	1.29	0.266	1.49	0.324	2.31	0.321	2.25	0.363	2.60
RatingB	0.463	2.33	0.414	2.06	0.455	2.90	0.450	2.80	0.514	3.34
# obs	229		172		171		171		167	
Adj R2	0.59		0.47		0.69		0.68		0.70	

NOTES: The dependent variable is the initial spread on the A notes. Estimation is by OLS. The omitted year is 1988. The omitted rating (of the sponsor) is AA; Rating B signifies Baa and Ba ratings. All regressions include trust dummies. For variable definitions, see the text.

which control for all average and time-invariant effects. The variation in a sponsor's rating over time is sufficient to cause significant changes over time in the yields paid by its ABS.

This result could be interpreted as suggesting that, even if the rating agencies place some weight on the risk of a sponsor in assessing the risk of their ABS notes, they do not do so fully. But the bond ratings are discretized, not continuous-valued, so there can be some differences in risk even among bonds with the same ratings. Also, investors' views of the risk might not completely coincide with the views of the ratings agencies. Hence we also directly control for the potential risk factors observable by investors. The next columns start by adding controls for the structure of the A notes. Of course, this structure is endogenous (but predetermined by the time of issuance) and should itself reflect the rating agencies' view of the notes' risk. Recall that the trust dummies already controlled for all time-invariant trust effects. These dummies are always jointly significant (unreported). For instance, some trusts might get locked into an older trust-structure technology that is considered riskier.

Column (2) explicitly controls for the amount of direct subordination behind each A note. LowSub is a dummy variable representing the quartile of notes with the smallest amount of subordination (i.e., the riskiest notes as measured by the relative size of their "buffer," *ceteris paribus)*. It has a significant positive coefficient. Thus, the notes with less enhancement have to offer investors higher yields to compensate. Nonetheless, the coefficients on the ratings variables change very little.[31] Column (3) adds as a control the expected maturity of the notes (Maturity). It also adds the size of the sellers' interest (SellersInt) and a dummy variable for whether the note is fixed rate or not (FixedRt). The results indicate that longer maturity and fixed-rate notes pay significantly higher spreads.[32] Given these controls the subordination measure (LowSub) becomes insignificant. This could mean that the size of the subordination might be a function of, among other things, maturity and whether the deal is fixed rate. Despite these effects, again the coefficients on the ratings do not change much. Column (4) controls for additional credit enhancement features, specifically dummy variables for the presence of a cash collateral account (I_CCA), a letter of credit (I_LOC), an internal reserve fund (I_RES), or other enhancement (I_Other). Given the other covariates, these additional enhancements are individually and jointly insignificant. (Though as indicated in Table 17.2, only CCAs are frequently used.) But the sponsor ratings remain significant.

31. Since LowSub is often missing, the sample size is smaller than in column (1). Nonetheless our conclusions below persist under the larger sample available if we do not control for LowSub.

32. Moody's (1995) noted a similar effect of maturity on spreads through 1993.

Finally, column (5) includes measures of the riskiness of the underlying portfolio of credit card receivables. Again, these are variables that the rating agencies take into account when approving the bond structure with a given rating, so their effects could already have been taken into account. The variable "Seasoned" is an indicator for older portfolios, with an average account age above 24 months. Since older accounts tend to have lower probabilities of default, this should reflect a safer portfolio.[33] Chargeoff is the initial (ex post) chargeoff rate in the portfolio.[34] Both variables are statistically significant, with the intuitive signs. Riskier portfolios, whether unseasoned or with higher chargeoff rates, must pay higher spreads. While Chargeoff is an ex post chargeoff rate, the conclusions are the same on instrumenting for it using the balance-weighted average chargeoff rate in the trust from the month before the issuance of each note in the sample. Even with these controls, the sponsor's rating remains significant.[35]

17.6.2. Analysis of the A-Note and B-Note Spreads

Table 17.4 repeats this analysis using both the A and B notes. All regressions now include an indicator variable (Junior) for the B notes. In column (1), this indicator is significantly positive, as expected given the greater risk of the B notes. They must pay on average 29 b.p. more than the A notes. The coefficient on the riskiest sponsors, RatingB, remains significant and large at 42 b.p. Thus the extra yield that must be paid by risky sponsors is even larger than the extra yield that must be paid by B notes. In column (2), LowSub indicates the A notes with the lowest quartile of subordination, and LowSubJr indicates the B notes with the lowest quartile of subordination. The latter variable is significant (and drives out

33. For an account-level analysis of the determinants of default probabilities, see Gross and Souleles (2002). For a portfolio-level analysis, see Musto and Souleles (2004). The original age data reflects the age of the accounts across the entire trust as of a given time. To estimate the age distribution of accounts underlying a given note at the time of issuance, we subtracted the time since closing. This assumes that the composition of the assets did not change too much between the time of closing and the time of reporting.

34. We take it from month three after issuance, since the excess spread components are sometimes missing in months one and two.

35. We also tried various extensions. For instance, we controlled for the importance of (on-balance sheet) credit card balances and other consumer receivables relative to total assets (CC/Assets). (When available from "Moody's Credit Opinions," CC/Assets is consumer receivables relative to assets. Otherwise, it is credit card balances relative to total assets from the *Call Report* data. In the latter case, in any given year CC/Assets is taken from the September quarter, and for 1988–90, it is taken from 1991:09.) CC/Assets had a significant negative effect on spreads, but did not change the results regarding the ratings. This suggests that the latter effect might not reflect just a correlation between the assets in the trust and the assets on-balance sheet, since presumably the credit card assets in the trust are more highly correlated with the credit card assets on-balance sheet, compared to other on-balance sheet assets.

Table 17-4. SPONSOR RATINGS AND INITIAL SPREADS ON A AND B NOTES

	(1)		(2)		(3)		(4)		(5)	
	coef.	**t**	**coef.**	**t**	**coef.**	**t**	**coef.**	**t**	**coef.**	**t**
Yr89	−0.565	−0.92	—		—		—		—	
Yr90	—		—		—		—		—	
Yr91	0.940	3.22	0.112	0.25	0.570	1.62	0.525	1.49	0.831	2.14
Yr92	0.922	2.39	0.937	1.16	1.303	2.06	1.292	2.04	1.251	1.88
Yr93	0.341	1.08	—		—		—		0.318	1.06
Yr94	0.264	0.89	−0.628	−2.68	−0.183	−0.99	−0.247	−1.31	0.472	1.94
Yr95	−0.770	−2.59	−1.382	−5.99	−0.965	−5.23	−1.024	−5.43	−0.356	−1.60
Yr96	−0.893	−3.04	−1.503	−6.49	−0.875	−4.68	−0.952	−4.92	−0.329	−1.57
Yr97	−0.891	−3.04	−1.508	−6.69	−0.946	−5.24	−1.010	−5.38	−0.406	−2.06
Yr98	−0.996	−3.35	−1.637	−7.24	−1.113	−6.20	−1.192	−6.38	−0.395	−2.29
Yr99	−0.727	−2.12	−1.411	−4.97	−0.919	−4.13	−1.000	−4.04	—	
LowSub			0.203	1.77	0.010	0.11	−0.023	−0.25	0.010	0.11
LowSubJr			0.350	2.66	0.096	0.92	0.066	0.62	0.116	1.10
Maturity					0.044	3.98	0.042	3.75	0.038	3.44
SellersInt					−0.032	−0.53	−0.022	−0.35	−0.010	−0.17

Table 17-4. (CONTINUED)

	(1)		(2)		(3)		(4)		(5)	
	coef.	**t**	**coef.**	**t**	**coef.**	**t**	**coef.**	**t**	**coef.**	**t**
FixedRt					0.858	13.22	0.878	13.05	0.889	13.70
I_CCA							−0.208	−1.65		
I_LOC							−0.250	−0.88		
I_RES							−0.271	−0.74		
I_Other							0.005	0.03		
Seasoned									−0.348	−3.86
Chargeoff									0.070	2.25
Junior	0.286	4.95	0.039	0.35	0.261	2.92	0.291	3.19	0.259	2.95
RatingA	0.154	1.15	0.215	1.56	0.285	2.66	0.274	2.54	0.331	3.01
RatingB	0.420	2.86	0.457	2.94	0.465	3.83	0.454	3.69	0.522	4.26
# obs	411		329		328		328		320	
Adj R2	0.63		0.52		0.72		0.72		0.72	

NOTES: See Table 17.3.

the direct effect of the Junior indicator), implying that B notes with less enhancement must pay higher yields. The rest of the analysis is analogous to that in Table 17.3, and the conclusions are the same.

Overall, the estimated effects of the sponsors' ratings appear to be robust. Even controlling for the ABS structure and underlying assets, the ratings of the sponsors remain significant, both statistically and economically. This supports our theoretical conclusion that the strength of the sponsor matters, because of the possibility of implicit recourse commitment. To reiterate, the trigger strategy at the root of the relational contract concerning recourse requires that the sponsor exist, that is, have not defaulted. The results are consistent with the investors in the ABS markets pricing the risk that the sponsor disappears and cannot support its SPVs.

17.7. EMPIRICAL TESTS: WHICH FIRMS SECURITIZE?

In this section we turn to testing whether riskier firms securitize more than others. Since our model is of course highly stylized we analyze more generally the determinants of securitization. We estimate equations of the following form, using the *Call Report* panel data from quarters 1991:09–2000:06:

$$\text{Securitize}_{i,t} = \beta_0' \mathbf{Time}_t + \beta_1' \mathbf{Bank}_i + \beta_2' X_{i,t} + \beta_3' \mathbf{Rating}_{i,t} + u_{i,t}, \qquad (17.3)$$

where $\text{Securitize}_{i,t}$ reflects the extent of credit-card securitization by bank i at time t, measured in one of three ways: i) We start with logit models of the probability that bank i has securitized, with dependent variable I_Sec being an indicator for whether the bank has any securitized credit card loans outstanding at time t (the extensive margin). ii) We also estimate Tobit models where the dependent variable Sec/Assets measures the amount of these securitizations normalized by total bank assets (including the securitized loans).[36] iii) To distinguish the intensive margin component in ii) from the extensive margin in i), we also estimate conditional OLS models of Sec/Assets conditional on Sec/Assets > 0.[37]

The dependent variables again include a full set of time dummies, this time quarter dummies. $\mathbf{X}_{i,t}$ controls for various bank characteristics over time. In particular it includes cubic polynomials in bank i's total assets, $\text{Assets}_{i,t}$, and in its

36. We include the securitized loans in assets in the denominator for convenience in interpreting Sec/Assets as a fraction ≤ 1. The denominator can also be interpreted as managed assets, although we do not have information on the full extent of off-balance sheet assets (including non-credit card assets) under management. Our conclusions are similar on when including the securitized loans in the denominator.

37. We would also like to estimate selection models, but we lack persuative omitted instrument.

share of credit card balances in total assets, $CC/Assets_{i,t}$. These control for scale effects, including costs that might arise in setting up and maintaining securitization trusts. We also control for the bank's capital ratio (equity capital divided by assets), $CapRatio_{i,t}$, again using a cubic polynomial.[38] Some specifications also control for all average and time-invariant bank effects ($\mathbf{Bank}_i$), using the corresponding fixed effects estimator. $\mathbf{Rating}_{i,t}$ is the Moody's rating of a bank's long-term senior obligations. Given the bank effects, the ratings variable will capture only within-bank variation, i.e., the effect of changes in a bank's rating over time on its propensity to securitize.[39]

Table 17.5 presents summary statistics for the key variables, for the entire sample period 1991–2000. To highlight the changes in the credit card ABS market over time, the second panel shows the same statistics for the end of the sample period (the first half of 2000). Comparing the panels shows the large growth in the market over the period. The fraction of banks that securitized (I_Sec) increased from about 8% in the early-to-mid 1990s to 15% at the end of the sample period, averaging about 11% overall during the period. The magnitude of securitizations relative to assets (Sec/Assets) increased from about 1.6% to 4.1% over the sample period, averaging 3.3%. The average bank rating declined over the sample period, though this happened for both the banks that securitized and those that did not.

Further, at any given time there is substantial cross-sectional variation across banks in the incidence and amount of securitization and in their ratings. The raw data suggest potential scale effects, with the big securitizers often being the bigger banks. These include highly rated securitizers, such as Citibank NV with an AA rating and Sec/Assets averaging about 71%. By contrast firms like Advanta (Sec/Assets $\approx$ 70%), Capital One ($\approx$ 57%), and Colonial ($\approx$ 65%) have lower ratings (RatingB). Given the potential problem of unobserved heterogeneity, our fixed effects estimators forego exploiting the purely cross-sectional average difference across banks; instead they set a high standard by relying on the more limited, but still substantial, within-bank variation over time in the incidence and amount of securitization and in the ratings. For instance, many banks were downgraded or upgraded at various times. Also, some banks securitized in only

38. We did not include the securitized loans (Sec) in assets in the denominator of CC/Assets or CapRatio, in order to avoid creating spurious correlations between these variables and the dependent variables (I_Sec and Sec/Assets). Calomiris and Mason (2004) discuss the relation between securitization and capital ratios.

39. The sample drops the few bank observations (about 10 banks) rated C and single B. Most of these were small banks in the early 1990s that did not securitize (only one of these banks securitized). As a result, they tended to be automatically dropped from the fixed effects estimation (or otherwise, their effect was imprecisely estimated due to their small sample size).

Table 17-5. SPONSOR RATINGS AND THE PROPENSITY TO SECURITIZE: SUMMARY STATISTICS

	1991–2000		2000	
	Mean	**s.d.**	**Mean**	**s.d.**
I_Sec	0.113	0.317	0.146	0.317
Sec/Assets	0.033	0.124	0.041	0.124
RatingAA	0.462	0.499	0.474	0.499
RatingA	0.446	0.497	0.397	0.497
RatingB	0.092	0.289	0.129	0.289
Assets (mil $)	16.0	39.1	25.4	39.1
CC/Assets	0.050	0.178	0.038	0.178
CapRatio	0.086	0.036	0.086	0.034
# obs	5012		363	

NOTES: In the first panel the sample is that for Table 17.6 columns (1) and (2), averaging over Call Report Data quarters 1991:09–2000:06. The second panel averages over only 2000:03 and 2000:06. See Table 17.6 and text for variable definitions.

a few years (perhaps just trying it out), whereas others securitized frequently but in varying amounts over time.

The main results are in Table 17.6. Column (1) begins with a logit model of the probability of securitizing (I_Sec), without bank effects. The effects of total assets (Assets), the importance of credit card assets (CC/Assets), and the capital ratio (CapRatio) are each jointly significant. Given the other covariates, in this specification the probability of securitizing is not monotonic in Assets; after initially increasing with Assets, it later declines. The probability of securitizing generally increases with CC/Assets (though declines a bit as CC/Assets gets very large). This could mean that having a large portfolio of credit cards provides economies of scale in securitizing. Also, the probability of securitizing is not monotonic in CapRatio (but increases for large CapRatio).

Of primary interest, at the bottom of the table, in this first specification the banks' ratings have a statistically significant, though non-monotonic, effect. Relative to the omitted AA ratings, the middle (RatingA) banks are somewhat less likely to securitize. Nonetheless, the riskiest (RatingB) are indeed much more likely to securitize.

Column (2) estimates a Tobit model of the amount of securitization (Sec/Assets). The conclusions are similar to those in the previous column. In both of these specifications, and those that follow, the pseudo and adjusted R2 statistics are relatively large.

The remaining columns control for bank fixed effects. Column (3) uses the fixed effects logit estimator. Note that as a result the sample size significantly declines, since this estimator drops banks for which I_Sec does not vary over time. Now the effect of Assets is monotonically increasing, though CC/Assets is less monotonic and CapRatio becomes insignificant. More importantly, both

Table 17-6. SPONSOR RATINGS AND THE USE OF SECURITIZATION

	(1) Logit			(2) Tobit			(3) Logit			(4) Cond. OLS		
	coef.	**s.e.**		**coef.**	**s.e.**		**coef.**	**s.e.**		**coef**	**se**	
Assets	0.031	0.004	**	0.006	0.001	**	0.235	0.039	**	−0.006	0.001	**
Assets2	−1.3E−04	2.5E−05	**	−2.8E−05	4.7E-06	**	-1.3E-03	2.7E−04	**	2.7E−05	6.3E−06	**
Assets3	1.4E−07	3.5E−08	**	3.1E−08	6.7E−09	**	2.4E−06	5.4E−07	**	−4.1E−08	1.2E−08	**
CC/Assets	5.092	2.393	**	0.891	0.411	**	53.172	11.598	**	0.095	0.203	
CC/Assets2	7.580	7.006		2.730	1.152	**	−110.737	29.759	**	0.736	0.507	
CC/Assets3	−9.369	5.049	*	−3.037	0.811	**	61.963	19.573	**	−0.926	0.338	**
CapRatio	21.53	7.46	**	5.46	1.35	**	18.82	31.39		2.99	1.39	**
CapRatio2	−91.93	36.87	**	−19.46	6.79	**	−142.06	133.26		−10.94	8.73	
CapRatio3	77.47	44.05	*	14.64	8.77	*	137.38	125.64		14.21	16.16	
RatingA	−0.552	0.120	**	−0.103	0.020	**	3.376	0.703	**	0.009	0.014	
RatingB	0.934	0.153	**	0.220	0.027	**	5.442	1.441	**	0.034	0.018	*
bank effects?	no			no			yes			yes		
# obs	5012			5012			730			568		
Pseudo/Adj R2	0.23			0.34								
Log-likelihood	−1369.0			−1083.5			−195.2			0.95		

NOTES: In columns (1) and (3), the dependent variable is the indicator I_Sec for whether the firm is currently securitizing (i.e., whether it has any securitized credit card loans currently outstanding). In column (2), it is the amount securitized normalized by assets (including the securitized loans), Sec/Assets. Column (3) uses the fixed effects logit estimator. In column (4), the dependent variable is Sec/Assets conditional on Sec/Assets>0. CC/Assets is credit card balances divided by assets. CapRatio is equity capital divided by assets. The omitted firm rating is AA. The sample includes the 1991:09–2000:06 Call Report Data, and all specifications include a complete set of quarter dummies.

Rating A and Rating B have significant positive effects, with a larger effect for the latter. Thus these results suggest that the probability of securitizing does indeed increase monotonically with banks' riskiness, consistent with our model. Column (4) instead focuses on the intensive margin, estimating a conditional OLS model of the fraction of securitized assets conditional on Sec/Assets > 0. CapRatio now has a monotonically increasing effect, though Assets has a negative effect on the intensive margin, and CC/Assets is not monotonic. While RatingA is positive but insignificant, RatingB has a larger positive coefficient, significant at the 6% level. Relative to banks with AA ratings, those with B ratings have about a 3.4 percentage point (p.p.) larger securitization fraction, on average. This is an economically significant effect, given that it is comparable in magnitude to the average Sec/Assets fraction of about 3.3 p.p.

Overall we conclude that there is some evidence that riskier firms are more likely to securitize, consistent with our model, though the effect is not always monotonic, depending on the specification. The effects of Assets, CC/Assets, and CapRatio are more sensitive to the specification.[40]

17.7.1. Summary

The empirical results are consistent with the theory proposed above, namely that an implicit contractual relationship between SPV sponsors and capital markets investors reduces bankruptcy costs. Consistent with the prediction that in the Implicit Recourse Equilibrium investors would price the risk of the sponsor defaulting, and hence being unable to subsidize the SPV, we found that the risk of the sponsor (as measured by the sponsor's bond rating) was consistently significant. The prediction of the model that firms with high expected bankruptcy costs would be the largest users of off-balance sheet financing was also generally confirmed.

17.8. CONCLUSION

Off-balance sheet financing is a pervasive phenomenon. It allows sponsoring firms to finance themselves by separating control rights over assets from financing. The operating entity, that is, the sponsoring firm, maintains control rights

40. We also tried various extensions. For instance, to see whether the ratings in turn might reflect the amount of securitization, we tried instrumenting for the ratings using lagged ratings. However it is not clear how long a lag would be best. At the extreme, we used the ratings from 1991:06, the quarter before the sample period starts. Given how small the credit card ABS market was at the time, it is unlikely that those ratings were significantly affected by securitization. The results were generally insignificant. This is not surprising, however, given the smaller sample size (since the 1991 ratings are not always available) and reduced amount of variation.

over the assets that generate cash flows. The assets (projects) can be financed by selling the cash flows to an SPV that has no need for control rights, because the cash flows have already been contracted for. We have argued that this arrangement is efficient because there is no need to absorb dead-weight bankruptcy costs with respect to cash flows that have already been contracted for. Off-balance sheet financing is about financing new projects by using cash flows promised under prior contracts as collateral. We showed that the efficient use of off-balance sheet financing is facilitated by an implicit arrangement, or contractual relations, between sponsoring firms and investors. The empirical tests, utilizing credit card asset-backed securitization as a testing ground, confirmed this interpretation of the SPV phenomenon.

APPENDIX: PROOFS

A. Lemma 1 Completion

It remains to verify that the equilibrium F derived under assumptions A3 and A4 is consistent. That is, we now restate assumptions A3 and A4 in terms of primitives. Recall A3 was stated as: $2y^L - h(e) < F$. The equilibrium F is given by:

$$F = \frac{D - (1 - e_H)^2 \left[2y^L (1 - c) - h(e_H)\right]}{e_H(2 - e_H)}.$$

Substituting the expression for F into A3 and simplifying gives:

$$2y^L \left[1 - c(1 - e_H)^2\right] - h(e_H) < D,$$

which is A3 stated in terms of primitives and consistent with the equilibrium.

Recall A4 was stated as: $2y^H - h(e) > y^H + y^L - h(e) > F$. Substitute the equilibrium value of F into $y^H + y^L - h(e) > F$, and simplify to obtain:

$$(e_H - 1)^2 y^L (1 - 2c) - h(e_H) > D. \; //$$

B. Solution to Problem (II)

Note that constraint (i) of Problem (II) in the main text can be written as:

$$e(2 - e)F^B + (1 - e)^2 \left[y^L (1 - c) - h(e)\right] \geq 0.5D.$$

Similarly, constraint (ii) of Problem (II) can be written as:

$$eF^S + (1 - e)y^L \geq 0.5D.$$

As before suppose lenders' beliefs are e_0. Then investors in the bank and SPV, respectively, will participate if the promised repayments are at least:

$$F_0^B = \frac{0.5D - (1-e_0)^2[y^L(1-c) - h(e_0)]}{e_0(2-e_0)},$$

and

$$F_0^S = \frac{0.5D - (1-e_0)y^L}{e_0}.$$

Substitute these into the bank's problem. Then the bank's problem is to choose $e \in \{e_H, e_L\}$ to:

$$\begin{aligned}\max V^S = {} & 2ey^H + e(1-e)y^L - e(2-e)h(e) \\ & -(1-\tau)e(2-e)\left[\frac{0.5D - (1-e_0)^2\left[y^L(1-c) - h(e_0)\right]}{e_0(2-e_0)}\right] \\ & -e\left[\frac{0.5D - (1-e_0)y^L}{e_0}\right]\end{aligned}$$

s.t. (iii) $V^S(e = e_H; e_0 = e_H) \geq V^S(e = e_L; e_0 = e_H)$ (Incentive Compatibility)

Suppose that beliefs are consistent, i.e., that $e = e_0 = e_H$. Then:

$$\begin{aligned}V^S = {} & 2e_Hy^H + e_H(1-e_H)y^L - e_H(2-e_H)h(e_H) \\ & -(1-\tau)[0.5D - (1-e_H)^2[y^L(1-c) - h(e_H)] \\ & -[0.5D - (1-e_H)y^L].\end{aligned} \tag{17.4}$$

LEMMA 2. If

$$\begin{aligned}2y^H(e_H - e_L) + {} & y^L[e_H(1-e_H) - e_L(1-e_L)] \\ & -h(e_H)e_H(2-e_H) + h(e_L)e_L(2-e_L) \\ & -(1-\tau)[0.5D - (1-e_H)^2[y^L(1-c) - h(e_H)] \\ & \left[1 - \frac{e_L(2-e_L)}{e_H(2-e_H)}\right] > 0\end{aligned}$$

then at the optimum, lenders believe $e_0 = e_H$ and the bank chooses $e = e_H$. The value of the bank V^S is given by (17.4).

Proof: The incentive compatibility constraint, $V^S(e = e_H;\ e_0 = e_H) \geq V^S(e = e_L;\ e_0 = e_H)$, is satisfied if the condition in the lemma holds. It remains to verify that the equilibrium F^B and F^S derived under A3a and A4a are consistent, i.e., to state A3a and A4a in terms of primitives. Recall A3a:

$2y^L - h(e) < F^B + F^S$. The equilibrium F^B and F^S are given by:

$$F^B = \frac{0.5D - (1-e_H)^2[y^L(1-c) - h(e_H)]}{e_H(2-e_H)},$$

and

$$F^S = \frac{0.5D - (1-e_H)y^L}{e_H}.$$

Substituting the expression for F^B and F^S into A3a and simplifying gives:

$$y^L(3-e_H) - h(e_H) + c(1-e_H)^2 y^L < 0.5D(3-e_H),$$

which is A3a stated in terms of primitives and consistent with the equilibrium.

Recall A4a: $2y^H - h(e) > y^H + y^L - h(e) > F^B + F^S$. Substitute the equilibrium values of F^B and F^S into $y^H + y^L - h(e) > F$, and simplify to obtain:

$$y^H e_H(2-e_H) + y^L(3 - 3e_H + e_H^2) - h(e_H) - cy^L(1-e_H)^2 > 0.5D(3-e_H)$$

which is A4a stated in terms of primitives and consistent with the equilibrium. //

C. Solution to Problem (III)

In solving Problem (III) we proceed as before and suppose lenders' beliefs are e_0. Then lenders will participate in lending to the bank and the SPV, respectively, if the promised repayments are at least:

$$F_0^C = \frac{0.5D - (1-e_0)^2[y^L(1-c) - h(e_0)}{e_0(2-e_0)},$$

and

$$F_0^{SC} = \frac{0.5D - (1-e_0)^2 y^L}{e_0(2-e_0)}.$$

Suppose that beliefs are consistent, i.e., $e = e_0 = e_H$. Then:

$$\begin{aligned} V^C = {} & 2e_H y^H + 2e_H(1-e_H)y^L - e_H(2-e_H)h(e_H) \\ & -(1-\tau)[0.5D - (1-e_H)^2[y^L(1-c) - h(e_H)] \\ & -[0.5D - (1-e_H)^2 y^L]] \end{aligned} \qquad \textbf{(17.5)}$$

LEMMA 3. If

$$2y^H(e_H - e_L) + 2y^L[e_H(1-e_H) - e_L(1-e_L)] - h(e_H)e_H(2-e_H) + h(e_L)e_L(2-e_L) - (1-\tau)[0.5D - (1-e_H)^2 [y^L(1-c) - h(e_H)]\left[1 - \frac{e_L(2-e_L)}{e_H(2-e_H)}\right] - [0.5D - (1-e_H)^2 y^L]\left[1 - \frac{e_L(2-e_L)}{e_H(2-e_H)}\right] > 0$$

then at the optimum, lenders believe $e_0 = e_H$ and the bank chooses $e_0 = e_H$. The value of the bank is given by (**17.5**).

Proof: The incentive compatibility constraint, $V^C(e = e_H;\ e_0 = e_H) \geq V^C(e = e_L;\ e_0 = e_H)$, is satisfied if the condition in the lemma holds. //

D. Proof of Proposition 3

Consider a bank that would choose securitization were it able to commit to subsidize its SPV in the state $\{y^H, y^L\}$, as in Problem III. Also, consider a date at which the bank has always subsidized its SPV in the past. Over the next period the bank is worth V^C if it securitizes one project off-balance sheet and retains the other on balance sheet. If both projects are financed on-balance sheet, the bank is worth V^H. By Propositions 1 and 2, $V^C > V^H$. The present value of this difference is the benefit to the bank of being able to utilize off-balance sheet financing, assuming that it continues to subsidize its SPV in the state $\{y^H,\ y^L\}$. Over the infinite horizon this annuity value is: $(V^C - V^H)/r$. (Recall that agents discount at rate r.)

At the end of the period, suppose that the state of the world is, in fact, $\{y^H, y^L\}$. Consider a one-shot deviation by the bank. That is, the bank decides not to subsidize the SPV, when investors expect the bank to subsidize it. From the expressions given above, the benefit to the bank of such a deviation is:

$$y^H - h(e_H) - (1-\tau)F^C > y^H + y^L - h(e_H) - (1-c)F^C - F^{SC}$$

which reduces to: $F^{SC} - y^L$.

To decide whether to deviate or not the bank compares the costs and benefits of deviation and chooses to subsidize the SPV as long as:

$$\frac{(V^C - V^H)}{r} > F^{SC} - y^L.$$

Substituting in this equation for V^C, V^H, and F^{SC} and simplifying (after some algebra) gives the quadratic inequality in the proposition. //

REFERENCES

Abreu, Dilip (1988), "On the Theory of Infinitely Repeated Games with Discounting," *Econometrica* 56, 383–396.

Ashman, Ian (2000), "Using Cayman Islands Special Purpose Vehicles," *International Financial Law Review* (April), 32–34.

Baker, George, Robert Gibbons, and Kevin Murphy (2002), "Relational Contracts and the Theory of the Firm," *Quarterly Journal of Economics* 117, 39–83.

Beatty, Anne, Philip Berger, and Joseph Magliolo (1995), "Motives for Forming Research and Development Financing Organizations," *Journal of Accounting and Economics* 19, 411–442.

Borgman, Richard and Mark Flannery (1997), "Loan Securitization and Agency: The Value of Originator-Provided Credit Enhancement," University of Florida, School of Business, working paper.

Calomiris, Charles and Joseph Mason (2004), "Credit Card Securitization and Regulatory Arbitrage," *Journal of Financial Services Research* 26, 5–28.

Croke, Jim (2003), "New Developments in Asset-Backed Commercial Paper," unpublished paper.

Elmer, Peter (1999), "Conduits: Their Structure and Risk," *FDIC Banking Review* 12, 27–40.

FitchIBCA (1999), "Implications of Securitization for Finance Companies," Financial Services *Special Report*, November 15, 1999.

FitchIBCA (2001), "Asset-Backed Commercial Paper Explained," *Structured Finance* (November 8, 2001).

Friedman, James W. (1971), "A Non-cooperative Equilibrium for Supergames," *Review of Economic Studies* 38, 1–12.

Gorton, Gary B. and George Pennacchi (1995), "Banks and Loan Sales: Marketing Non-Marketable Assets," *Journal of Monetary Economics* 35(3), 389–411.

Gorton, Gary B. and George Pennacchi (1989) "Are Loan Sales Really Off-Balance Sheet?," *Journal of Accounting, Auditing and Finance* 4:2, 125–45.

Gorton, Gary B. and Andrew Winton (2003). "Financial Intermediation," in *The Handbook of the Economics of Finance: Corporate Finance*, edited by George Constantinides, Milton Harris, and Rene Stulz (Elsevier Science; 2003) (NBER Working Paper # 8928).

Green, Edward and Robert H. Porter (1984), "Noncooperative Collusion under Imperfect Price Information," *Econometrica* 52, 87–100.

Gross, David and Nicholas S. Souleles (2002). "An Empirical Analysis of Personal Bankruptcy and Delinquency," *Review of Financial Studies,* 15(1), 319–347.

Henry, David, Heather Timmons, Steve Rosenbush, and Michael Arndt (2002), "Who else is hiding debt?," *Business Week* (January 28), 36–37.

Higgins, Eric and Joseph Mason (2004), "What is the Value of Recourse to Asset Backed Securities? A Study of Credit Card Bank ABS Rescues," *Journal of Banking and Finance* 28, 857–874.

Hodge, J.B. (1996), "The Use of Synthetic Leases to Finance Build-to-Suit Transactions," *Real Estate Finance Journal* 11, 17–21.

Hodge, J.B. (1998), "The Synthetic Lease: Off-Balance-Sheet Financing of the Acquisition of Real Property," *Real Estate Finance Journal* 14, 159–76.

Humphreys, Thomas and R.M. Kreistman (1995), *Mortgage-Backed Securities including REMICs and Other Investment Vehicles* (New York; Little, Brown).
Johnson, Kathleen (2002), "Consumer Loan Securitization," in Durkin, Thomas A. and Michael E. Staten (2002), *The Impact of Public Policy on Consumer Credit* (Boston; Kluwer Academic Publishers).
Kendall, Leon T. and Michael J. Fishman (1996), *A Primer on Securitization* (Cambridge, MA; MIT Press).
Klee, Kenneth and Brendt Butler (2002), "Asset-Backed Securitization, Special Purpose Vehicles and Other Securitization Issues," *Uniform Commercial Code Law Journal* 35, 23–67.
Kramer, Andrea (2003), *Financial Products: Taxation, Regulation and Design*, 3 volumes (Aspen Publishers; New York City).
Langbein, John H. (1997), "The Secret Life of the Trust: The Trust as an Instrument of Commerce," *Yale Law Journal* 107, 165–189.
Lim, Steve, Steve Mann, and Vassil Mihov (2003), "Market Evaluation of Off-Balance Sheet Financing: You Can Run But You Can't Hide," Texas Christian University, working paper.
Mills, Lillian and Kaye Newberry (2004), "Firms' Off-Balance Sheet Financing: Evidence from their Book-Tax Reporting Differences," University of Arizona working paper.
Moody's Investors Service (May 29, 2003), "Securitization in New Markets: Moody's Perspective: Europe, Africa and the Middle East," International Structured Finance, *Special Report.*
Moody's Investors Service (2002), "Securitization and its Effect on the Credit Strength of Companies: Moody's Perspective 1987–2002," *Special Comments.*
Moody's Investors Service (August 30, 2002), "Bullet Proof Structures Revisited: Bankruptcies and a Market Hangover Test Securitizations' Mettle," *Special Report.*
Moody's Investors Service (September 1997), "Alternative Financial Ratios for the Effects of Securitization," in "Securitization and its Effect on the Credit Strength of Companies: Moody's Perspective 1987–2002," *Special Comments* (2002).
Moody's Investors Service (January 1997), "The Costs and Benefits of Supporting 'Troubled' Asset-Backed Securities: Has the Balance Shifted?," in "Securitization and its Effect on the Credit Strength of Companies: Moody's Perspective 1987–2002," *Special Comments* (2002).
Moody's Investors Service (May 1995), "Spread Thin: An Empirical Investigation of Yields on Credit Card Asset-Backed Securities," *Special Report.*
Moody's Investors Service (November 11, 1994), "The 'C' Tranches of Credit Card-Backed Securities: Credit Risks for Investors Vary," Structured Finance, *Special Report.*
Moody's Investors Service (April 1993), "Asset-Backed Commercial Paper: Understanding the Risks," *Special Report.*
Musto, David, and Nicholas S. Souleles (2004), "A Portfolio View of Consumer Credit," University of Pennsylvania working paper.
Office of the Comptroller of the Currency, Federal Deposit Insurance Corporation, Board of Governors of the Federal Reserve System, and the Office of Thrift Supervision (2002), "Interagency Guidance on Implicit Recourse in Asset Securitizations," (May 23, 2002).

Peaslee, J. and D. Nirenberg (2001), *Federal Income Taxation of Securitization Transactions* (3rd. ed.; Frank J. Fabozzi Associates).

Pfister, Benedicte (2000), "Whole Business Securitizations: A Unique Opportunity for UK Assets," *International Structured Finance Special Report*, Moody's Investors Service (October 19, 2000).

Restatement (Third) of the Law, Trusts, Volumes 1 and 2 (American Law Institute; 2003).

Rotemberg, Julio and Garth Saloner (1986), "A Supergame-Theoretic Model of Price Wars During Booms," *American Economic Review* 76, 390–407.

Schwarcz, Steven (2003a), *Structured Finance*, third edition (Practicing Law Institute; New York City).

Schwarcz, Steven (2003b), "Commercial Trusts as Business Organizations: Unraveling the Mystery," *The Business Lawyer* 58 (February), 559–585.

Shakespeare, Catherine (2003), "Do Managers use Securitization Volume and Fair Value Estimates to Hit Earnings Targets?," University of Michigan, School of Business, working paper.

Shakespeare, Catherine (2001), "Accounting for Asset Securitizations: Complex Fair Values and Earnings Management," University of Michigan, School of Business, working paper.

Shevlin, Terrence (1987), "Taxes and Off-Balance Sheet Financing: Research and Development Limited Partnerships," *The Accounting Review* 52, 480–509.

Sitkoff, Robert H. (2003), "Trust Law, Corporate Law, and Capital Market Efficiency," University of Michigan Law School, John M. Olin Center for Law & Economics Working Paper No. 20.

Standard and Poor's (no date), *Structured Finance: Credit Card Criteria.*

Standard and Poor's (2002), *U.S. Legal Criteria for "Recycled" Special Purpose Entities.*

Weidner, Donald (2000), "Synthetic Leases: Structure Finance, Financial Accounting and Tax Ownership," Florida State University, College of Law, Working Paper No. 06 (April 2000).

PART V

The Crisis of 2007–2008

18

Questions and Answers about the Financial Crisis*

GARY B. GORTON ■

> Unfortunately the subject [of the Panic of 1837] has been connected with the party politics of the day. Nothing can be more unfavorable to the development of truth, on questions in political economy, than such a connection. A good deal which is false, with some admixture of truth, has been put forward by political partisans on either side. As it is the wish of the writer that the subject should be discussed on its own merits and free from such contaminating connection, he has avoided as much as possible all reference to the political parties of the day (Appleton (1857), May 1841).

> The current explanations [of the Panic of 1907] can be divided into two categories. Of these the first includes what might be called the superficial theories. Thus it is commonly stated that the outbreak of a crisis is due to a lack of confidence—as if the lack of confidence was not itself the very thing which needs to be explained. Of still slighter value is the attempt to associate a crisis with some particular governmental policy, or with some action of a country's executive. Such puerile interpretations have commonly been confined to countries like the United States where the political passions of a democracy had the fullest sway. . . . Opposed to these popular, but wholly unfounded, interpretations is the second class of explanations, which seek to burrow beneath the surface and to discover the more . . . fundamental causes of the periodicity of crises (Seligman (1908), p. xi).

* Thanks to Lori Gorton, Stephen Partridge-Hicks, Andrew Metrick, and Nick Sossidis for comments and suggestions.

> The subject [of the Panic of 1907] is technical. Opinions formed without a grasp of the fundamental principles and conditions are without value. The verdict of the uninformed majority gives no promise of being correct . . . . If to secure proper banking legislation now it is necessary for a . . . campaign of public education, it is time it were begun (Vanderlip (1908), p. 18).

> Don't bother me with facts, son. I've already made up my mind.
>
> —FOGHORN LEGHORN

18.1. INTRODUCTION

Yes, we have been through this before, tragically many times.

U.S. financial history is replete with banking crises and the predictable political responses. Most people are unaware of this history, which we are repeating. A basic point of this note is that there is a fundamental, structural, feature of banking, which if not guarded against leads to such crises. Banks create money, which allows the holder to withdraw cash on demand. The problem is not that we have banking; we need banks and banking. And we need this type of bank product. But, as the world grows and changes, this money feature of banking reappears in different forms. The current crisis, far from being unique, is another manifestation of this problem. The problem then is structural.

In this note, I pose and try to answer what I think are the most relevant questions about the crisis. I focus on the systemic crisis, not other attendant issues. I do not have all the answers by any means. But, I know enough to see that the level of public discourse is politically motivated and based on a lack of understanding, as it has been in the past, as the opening quotations indicate. The goal of this note is to help raise the level of discourse.

18.2. QUESTIONS AND ANSWERS

Q. What happened?

A. This question, though the most basic and fundamental of all, seems very difficult for most people to answer. They can point to the effects of the crisis, namely the failures of some large firms and the rescues of others. People can point to the amounts of money invested by the government in keeping some firms running. But they can't explain what actually happened, what caused these firms to get into trouble. Where and how were losses actually realized? What actually happened? The remainder of this short note will address these questions. I start with an overview.

There was a banking panic, starting August 9, 2007. In a banking panic, depositors rush en masse to their banks and demand their money back. The banking system cannot possibly honor these demands because they have lent the money out or they are holding long-term bonds. To honor the demands of depositors, banks must sell assets. But only the Federal Reserve is large enough to be a significant buyer of assets.

Banking *means* creating short-term trading or transaction securities backed by longer term assets. Checking accounts (demand deposits) are the leading example of such securities. The fundamental business of banking creates a vulnerability to panic because the banks' trading securities are short term and need not be renewed; depositors can withdraw their money. But, panic can be prevented with intelligent policies. What happened in August 2007 involved a different form of bank liability, one unfamiliar to regulators. Regulators and academics were not aware of the size or vulnerability of the new bank liabilities.

In fact, the bank liabilities that we will focus on are actually very old, but have not been quantitatively important historically. The liabilities of interest are sale and repurchase agreements, called the "repo" market. Before the crisis trillions of dollars were traded in the repo market. The market was a very liquid market like another very liquid market, the one where goods are exchanged for checks (demand deposits). Repo and checks are both forms of money. (This is not a controversial statement.) There have always been difficulties creating private money (like demand deposits) and this time around was no different.

The panic in 2007 was not observed by anyone other than those trading or otherwise involved in the capital markets because the repo market does not involve regular people, but firms and institutional investors. So, the panic in 2007 was not like the previous panics in American history (like the Panics of 1837, 1857, 1873, 1893, 1907, and so on) in that it was not a mass run on banks by individual depositors, but instead was a run by firms and institutional investors on financial firms. The fact that the run was not observed by regulators, politicians, the media, or ordinary Americans has made the events particularly hard to understand. It has opened the door to spurious, superficial, and politically expedient "explanations" as well as demagoguery.

Q. How could there be a banking panic when we have deposit insurance?

A. As explained, the Panic of 2007 was not centered on demand deposits, but on the repo market which is not insured.

As the economy transforms with growth, banking also changes. But, at a deep level the basic form of the bank liability has the same structure, whether it is private bank notes (issued before the Civil War), demand deposits, or sale and repurchase agreements. Bank liabilities are designed to be safe; they are short term, redeemable, and backed by collateral. But, they have always been vulnerable to mass withdrawals, a panic. This time the panic was in the sale and

repurchase market ("repo market"). But, before we come to that we need to think about how banking has changed.

Americans frequently experienced banking panics from colonial days until deposit insurance was passed in 1933, effective 1934. Government deposit insurance finally ended the panics that were due to demand deposits (checking accounts). A demand deposit allows you to keep money safely at a bank and get it any time you want by asking for your currency back. The idea that you can redeem your deposits anytime you want is one of the essential features of making bank debt safe. Other features are that the bank debt is backed by sufficient collateral in the form of bank assets.

Before the Civil War the dominant form of money was privately issued bank notes; there was no government currency issued. Individual banks issued their own currencies. During the Free Banking Era, 1837–1863, these currencies had to be backed by state bonds deposited with the authorities of whatever state the bank was chartered in. Bank notes were also redeemable on demand and there were banking panics because sometimes the collateral (the state bonds) was of questionable value. This problem of collateral will reappear in 2007.

During the Free Banking Era banking slowly changed, first in the cities, and over the decades after the Civil War nationally. The change was that demand deposits came to be a very important form of bank money. During the Civil War the government took over the money business; national bank notes ("greenbacks") were backed by U.S. Treasury bonds and there were no longer private bank notes. But, banking panics continued. They continued because demand deposits were vulnerable to panics. Economists and regulators did not figure this out for decades. In fact, when panics due to demand deposits were ended it was not due to the insight of economists, politicians, or regulators. Deposit insurance was not proposed by President Roosevelt; in fact, he opposed it. Bankers opposed it. Economists decried the "moral hazards" that would result from such a policy. Deposit insurance was a populist demand. People wanted the dominant medium of exchange protected. It is not an exaggeration to say that the quiet period in banking from 1934 to 2007, due to deposit insurance, was basically an accident of history.

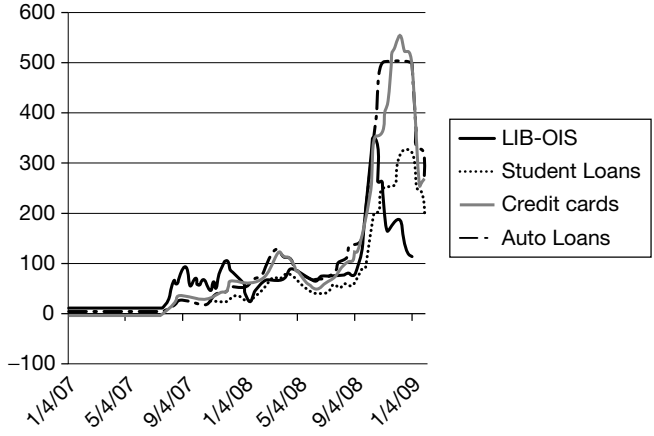
Times change. Now, banking has changed again. In the last 25 years or so, there has been another significant change: a change in the form and quantity of bank liabilities that has resulted in a panic. This change involves the combination of securitization with the repo market. At root this change has to do with the traditional banking system becoming unprofitable in the 1980s. During that decade, traditional banks lost market share to money market mutual funds (which replaced demand deposits) and junk bonds (which took market share from lending), to name the two most important changes. Keeping passive cash flows on the balance sheet from loans, when the credit decision was already

made, became unprofitable. This led to securitization, which is the process by which such cash flows are sold. I discuss securitization below.

Q. What has to be explained to explain the crisis?

A. It is very important to set standards for the discussion. I think we should insist on three criteria.

First, a coherent answer to the question of what happened must explain why the spreads on asset classes completely unrelated to subprime mortgages rose dramatically. (Or, to say it another way, the prices of bonds completely unrelated to subprime fell dramatically.) The figure below shows the LIBOR-OIS spread, a measure of interbank counterparty risk, together with the spreads on AAA tranches of bonds backed by student loans, credit card receivables, and auto loans. The units on the y-axis are basis points (a "basis point" is 1/100 of a percentage point). The three types of bonds normally trade near or below LIBOR. Yet, in the crisis, they spiked dramatically upwards and they moved with the measure of bank counterparty risk. Why?



SOURCE: Gorton and Metrick (2009a).

The outstanding amount of subprime bonds was not large enough to cause a systemic financial crisis by itself. It does not explain the figure above. No popular theory (academic or otherwise) explains the above figure. Let me repeat that another way. Common "explanations" are too vague and general to be of any value. They do not explain what actually happened. The issue is why **all** bond prices plummeted. What caused that?

This does not mean that there are not other issues that should be explored, as a matter of public policy. Nor does it mean that these other issues are not important. It does, however, mean that these other issues—whatever they are—are irrelevant to understanding the main event of the crisis.

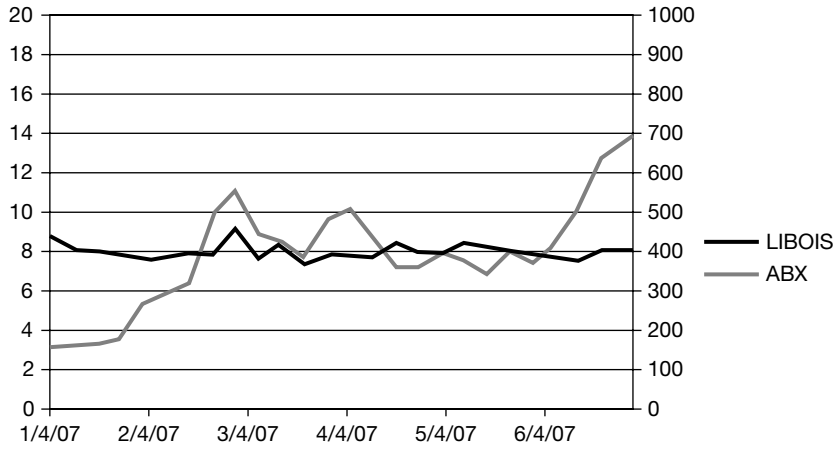
Second, an explanation should be able to show exactly how losses occurred. This is a different question than the first question. Prices may go down, but how did that result in trillions of dollars of losses for financial firms?

Finally, a convincing answer to the question of what happened must include some evidence and not just be a series of broad, vague, assertions.

In what follows I will try to adhere to these criteria.

Q. Wasn't the panic due to subprime mortgages going bad due to house prices falling?

A. No. This cannot be the whole story. Outstanding subprime securitization was not large enough by itself to have caused the losses that were experienced. Further, the timing is wrong. Subprime mortgages started to deteriorate in January 2007, eight months before the panic in August. The gray line below is the BBB tranche of the ABX index, a measure of subprime fundamentals. It is in the form of a spread, so when it rises it means that the fundamentals are deteriorating. The two axes are measured in basis points; the axis on the right side is for the ABX. The black line, the one that is essentially flat, is the LIBOR minus OIS spread—a measure of counterparty risk in the banking system. It is measured on the left-hand axis. The point is this: Subprime started significantly deteriorating well before the panic, which is not shown here. Moreover, subprime was never large enough to be an issue for the global banking system. In 2007 subprime stood at about $1.2 trillion outstanding, of which roughly 82 percent was rated AAA and to date has very small amounts of realized losses. Yes, $1.2 trillion is a large number, but for comparison, the total size of the traditional and parallel banking systems is about $20 trillion.



SOURCE: Gorton and Metrick (2009a). LIBOIS is the LIBOR minus Overnight Index Swap spread. ABX refers to the spread on the BBB tranche of the ABX index.

Subprime will play an important role in the story later. But by itself it does not explain the crisis.

Q. Subprime mortgages were securitized. Isn't securitization bad because it allows banks to sell loans?

A. Holding loans on the balance sheets of banks is not profitable. This is a fundamental point. This is why the parallel or shadow banking system developed. If an industry is not profitable, the owners exit the industry by not investing; they invest elsewhere. Regulators can make banks do things, like hold more capital, but they cannot prevent exit if banking is not profitable. "Exit" means that the regulated banking sector shrinks, as bank equity holders refuse to invest more equity. Bank regulation determines the size of the regulated banking sector, and that is all. One form of exit is for banks to not hold loans but to sell the loans; securitization is the selling of portfolios of loans. Selling loans—while news to some people—has been going on now for about 30 years without problems.

In securitization, the bank is still at risk because the bank keeps the residual or equity portion of the securitized loans and earns fees for servicing these loans. Moreover, banks support their securitizations when there are problems. No one has produced evidence of any problems with securitization generally; though there are have been many such assertions. The motivation for banks to sell loans is profitability. In a capitalist economy, firms (including banks) make decisions to maximize profits. Over the last 25 years securitization was one such outcome. As mentioned, regulators cannot make firms do unprofitable things because investors do not have to invest in banks. Banks will simply shrink. This is exactly what happened. The traditional banking sector shrank, and a whole new banking sector developed—the outcome of millions of individual decisions over a quarter of a century.

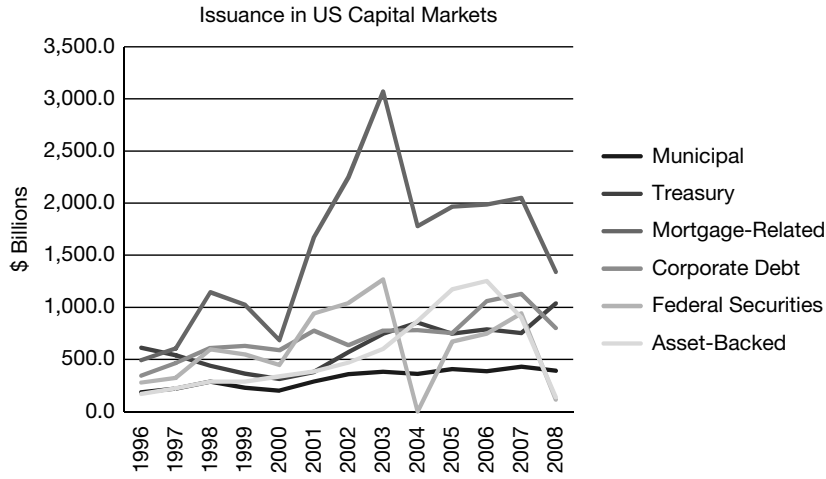
Q. What is this new banking system, the "parallel banking system" or "shadow banking system" or "securitized banking system"?

A. A major part of it is securitization. Never mind the details for our present purposes (see Gorton (2010) for details); the main point is that this market is very large. The figure below shows the issuance amounts of various levels of fixed-income instruments in the capital markets. The mortgage-related instruments, including securitization is the largest market.

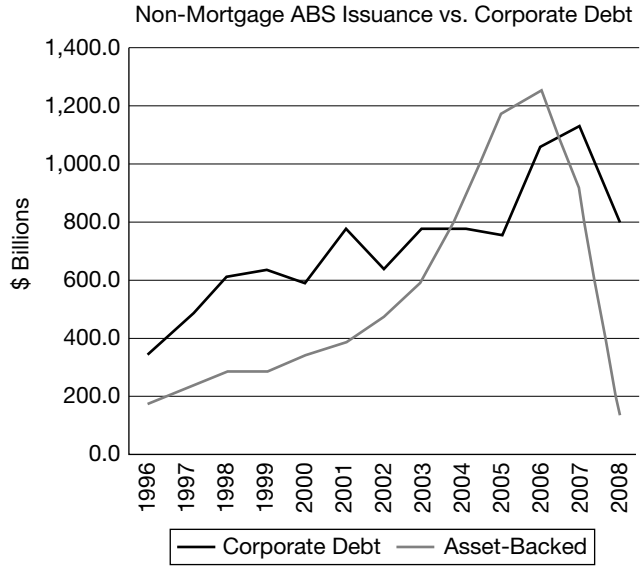
Of greater interest perhaps is the comparison of the non-mortgage securitization (labeled "Asset-Backed" in the above figure) issuance amounts with the amount of all of U.S. corporate debt issuance.

This is portrayed in the figure below.

The figure shows two very important points. First, measured by issuance, *non-mortgage* securitization exceeded the issuance of all U.S. corporate debt starting



SOURCES: U.S. Department of Treasury, Federal Agencies, Thomson Financial, Inside MBS & ABS, Bloomberg.

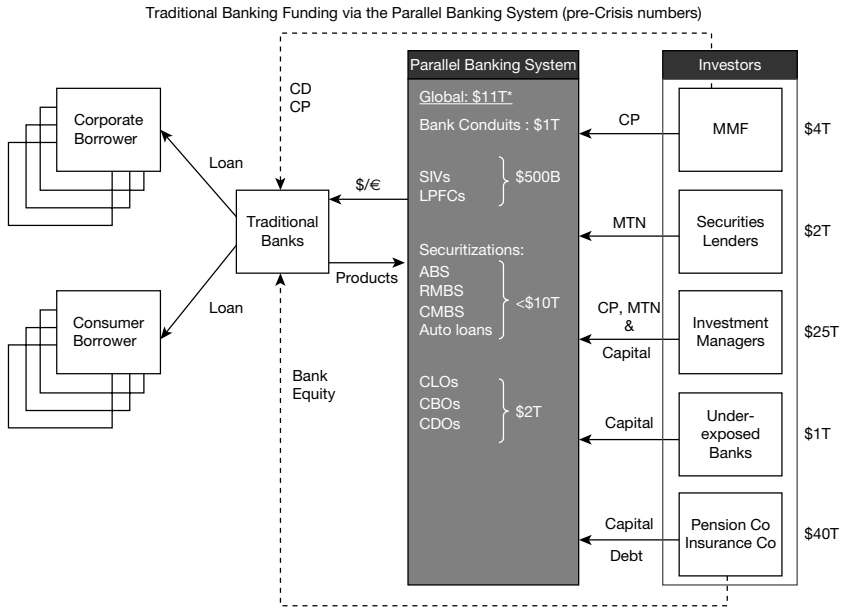


SOURCES: U.S. Department of Treasury, Federal Agencies, Thomson Financial, Inside MBS & ABS, Bloomberg.

in 2004. Secondly, the figure shows the effects of the crisis on issuance: this market is essentially dead.

Q. So, traditional, regulated, banks sell their loans to the other banking system. Is that the connection between the parallel or shadow banking system and the traditional banking system?

Traditional Banking Funding via the Parallel Banking System (pre-Crisis numbers)



SOURCE: Gordian Knot.

A. Yes. The parallel or shadow banking system is essentially how the traditional, regulated, banking system is funded. The two banking systems are intimately connected. This is very important to recognize. It means that without the securitization markets the traditional banking system is not going to function. The diagram above shows how the two banking systems are related.

The figure shows how the traditional banking system funded its activities just prior to the crisis. The loans made to consumers and corporations, on the left side of the figure, correspond to the credit creation that the traditional banks are involved in. Where do they get the money to lend to corporations and consumers? Portfolios of the loans are sold as bonds, to the various securitization vehicles in the parallel banking system (the gray box in the middle). These vehicles are securitization, conduits, structured investment vehicles (SIVs), limited purpose finance corporations (LPFCs), collateralized loan obligations (CLOs), collateralized bond obligations (CBOs), collateralized debt obligations (CDOs), and specialist credit managers. Like the traditional banks, these vehicles are intermediaries. They in turn are financed by the investors on the right side of the figure.

Q. But weren't these securitizations supposed to be distributed to investors? Why did banks keep so much of this on their balance sheets?

A. Above we discussed the reasons that securitization arose, the supply of securitized products. What about the demand? There is a story that is popularly called

"originate-to-distribute" which claims that securitizations should not end up on bank balance sheets. There is no basis for this idea. In fact, there is an important reason for why banks did hold some of these bonds: these bonds were needed as collateral for a form of depository banking. The other part of the new banking sector involves the new "depositors." This part of the story is not shown in the figure above.

Institutional investors and nonfinancial firms have demands for checking accounts just like you and I do. But, for them there is no safe banking account because deposit insurance is limited. So, where does an institutional investor go to deposit money? The institutional investor wants to earn interest, have immediate access to the money, and be assured that the deposit is safe. But, there is no checking account insured by the FDIC if you want to deposit $100 million. Where can this depositor go?

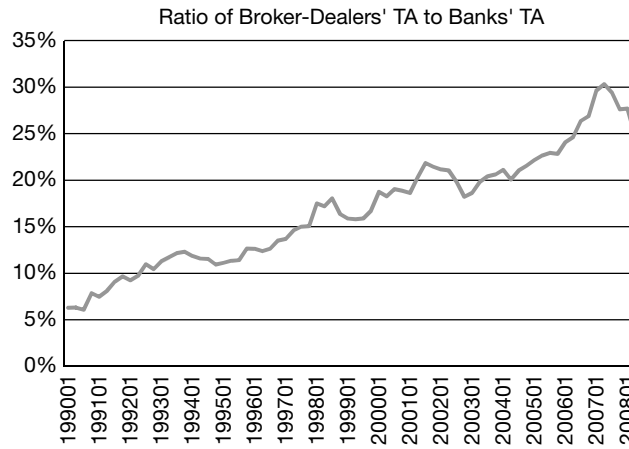
The answer is that the institutional investor goes to the repo market. For concreteness, let's use some names. Suppose the institutional investor is Fidelity, and Fidelity has $500 million in cash that will be used to buy securities, but not right now. Right now Fidelity wants a safe place to earn interest, but such that the money is available in case the opportunity for buying securities arises. Fidelity goes to Bear Stearns and "deposits" the $500 million *overnight* for interest. What makes this deposit safe? The safety comes from the collateral that Bear Stearns provides. Bear Stearns holds some asset-backed securities that are earning 6 percent. They have a market value of $500 million. These bonds are provided to Fidelity as collateral. Fidelity takes physical possession of these bonds. Since the transaction is overnight, Fidelity can get its money back the next morning, or it can agree to "roll" the trade. Fidelity earns, say, 3 percent.

Just like banking throughout history, Bear has, for example, borrowed at 3 percent and "lent" at 6 percent. In order to conduct this banking business Bear needs collateral (that earns 6 percent in the example)—just like in the Free Banking Era banks needed state bonds as collateral. In the last 25 years or so money under management in pension funds and institutional investors, and money in corporate treasuries, has grown enormously, creating a demand for this kind of depository banking.

How big was the repo market? No one knows. The Federal Reserve only measures repo done by the 19 primary dealer banks that it is willing to trade with. So, the overall size of the market is not known. I roughly guess that it is at least $12 trillion, the size of the total assets in the regulated banking sector. The fact is, however, that the repo market was never properly measured, so we will likely never know for sure how big it was. There is indirect evidence, however, that we can we bring to bear on this question.

One thing we can look at is how big the broker-dealer banks were compared to the traditional banks. Broker-dealer banks to a large extent were the new depository institutions. Since repo requires collateral, thee banks would need to

grow their balance sheets to hold the collateral needed for repo. Broker-dealers are essentially the old investment banks. While this division is not strictly correct, it gives some idea. The figure below shows the ratio of the total assets of broker-dealers to total assets of the regulated banks.



SOURCE: Flow of Funds data; Gorton and Metrick (2009a).

You can see in the figure that the ratio of total assets of broker-dealer banks to traditional banks was about 6 percent in 1990, and had grown to about 30 percent just before the crisis onset. In the meantime, as we saw above, securitization was growing enormously over the same period. Why would dealer banks be growing their balance sheets if there was not some profitable reason for this? My answer is that the new depository business using repo was also growing.

Now, of course there is the alterative hypothesis, that the broker-dealer banks were just irresponsible risk-takers. They held all these long-term assets financing them with short-term repo just to take on risk. (Of course there are much easier ways to take on (much more) risk.) As a theory of the crisis this "theory" is hard to understand. It is a lazy "explanation" in the form of Monday morning quarterbacking. Further, this view, of course, ignores the fact that someone must be on the other side of the repo. Who were the depositors? What was their incentive to engage in this if it was just reckless bankers?

Q. Why doesn't the repo market just use Treasury bonds for collateral?

A. A problem with the new banking system is that it depends on collateral to guarantee the safety of the deposits. But, there are many demands for such collateral. Foreign governments and investors have significant demands for U.S. Treasury bonds, U.S. agency bonds, and corporate bonds (about 40 percent is held by foreigners). Treasury and agency bonds are also needed to collateralize derivatives positions. Further, they are needed to use as collateral for clearing

and settlement of financial transactions. There are few AAA corporate bonds. Roughly speaking (which is the best that can be done, given the data available), the total amount of possible collateral in U.S. bond markets, minus the amount held by foreigners is about $16 trillion. The amount used to collateralize derivatives positions (according to ISDA) is about $4 trillion. It is not known how much is needed for clearing and settlement. Repo needs, say, $12 trillion.

The demand for collateral has been largely met by securitization, a 30-year old innovation that allows for efficient financing of loans. Repo is to a significant degree based on securitized bonds as collateral, a combination called "securitized banking." The shortage of collateral for repo, derivatives, and clearing/settlement is reminiscent of the shortages of money in early America, which is what led to demand deposit banking.

Q. Ok, let's assume that the repo market is very large. You say the events were a "panic," how do we know this is so? What does this have to do with repo?

A. Here's where we come to the question of "what happened."

There's another aspect to repo that is important: haircuts. In the repo example I gave above, Fidelity deposited $500 million of cash with Bear Stearns and received as collateral $500 million of bonds, valued at market value. Fidelity does not care if Bear Stearns becomes insolvent because Fidelity in that event can unilaterally terminate the transaction and sell the bonds to get the $500 million. That is, repo is not subject to Chapter 11 bankruptcy; it is excluded from this.

Imagine that Fidelity said to Bear: "I will deposit only $400 million and I want $500 million (market value) of bonds as collateral." This would be a 20 percent haircut. In this case Fidelity is protected against a $100 million decline in the value of the bonds, should Bear become insolvent and Fidelity want to sell the bonds.

Note that a haircut requires the bank to raise money. In the above example, suppose the haircut was zero to start with, but then it becomes positive, say that it rises to 20 percent. This is essentially a withdrawal from the bank of $100 million. Bear turns over $500 million of bonds to Fidelity, but only receives $400 million. This is a withdrawal of $100 million from the bank. How does Bear Stearns finance the other $100 million? Where does the money come from? We will come to this shortly.

Prior to the panic, haircuts on all assets were zero for high quality dealer banks!

For now, keep in mind that an increase in the haircuts is a withdrawal from the bank. Massive withdrawals are a banking panic. That's what happened. Like during the pre-Federal Reserve panics, there was a shock that by itself was not large, house prices fell. But, the distribution of the risks (where the subprime bonds were, in which firms, and how much) was not known. Here is where subprime plays its role. Elsewhere, I have likened subprime to e-coli (see Gorton (2009a,

2010)). Millions of pounds of beef might be recalled because the location of a small amount of e-coli is not known for sure. If the government did not know which ground beef possibly contained the e-coli, there would be a panic: people would stop eating ground beef. If we all stop eating hamburgers for a month, or a year, it would be a big problem for McDonald's, Burger King, Wendy's and so on. They would go bankrupt. That's what happened.

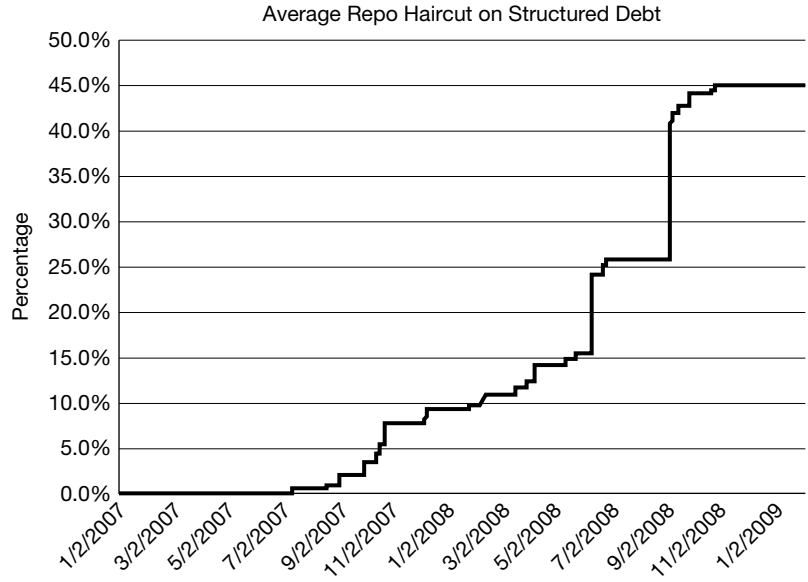
The evidence is in the figure below, which shows the increase in haircuts for securitized bonds (and other structured bonds) starting in August 2007.

The figure is a picture of the banking panic. We don't know how much was withdrawn because we don't know the actual size of the repo market. But, to get a sense of the magnitudes, suppose the repo market was $12 trillion and that repo haircuts rose from zero to an average of 20 percent. Then the banking system would need to come up with $2 trillion, an impossible task.

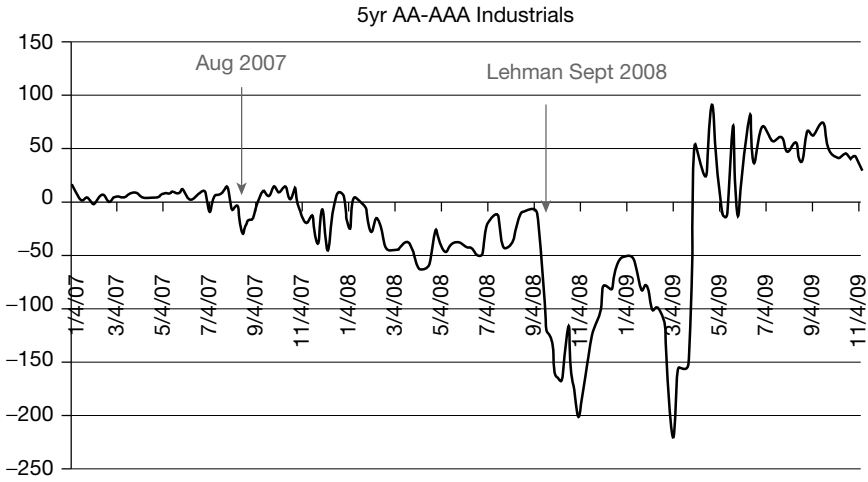
Q. Where did the losses come from?

A. Faced with the task of raising money to meet the withdrawals, firms had to sell assets. They were no investors willing to make sufficiently large new investments, on the order of $2 trillion. In order to minimize losses firms chose to sell bonds that they thought would not drop in price a great deal, bonds that were not securitized bonds, and bonds that were highly rated. For example, they sold Aaa-rated corporate bonds.

These kinds of forced sales are called "fire sales"—sales that must be made to raise money, even if the sale causes to price to fall because so much is offered for



SOURCE: Gorton and Metrick (2009a).



SOURCE: Gorton and Metrick (2009a).

sale, and the seller has no choice but to take the low price. The low price reflects the distressed, forced, sale, not the underlying fundamentals. There is evidence of this. Here is one example. Normally, Aaa-rated corporate bonds would trade at higher prices (lower spreads) than, say, Aa-rated bonds. In other words, these bonds would fetch the most money when sold. However, when all firms reason this way, it doesn't turn out so nicely.

The figure above shows the spread between Aa-rated corporate bonds and Aaa-rated corporate bonds, both with five year maturities. This spread should always be positive, unless so many Aaa-rated corporate bonds are sold that the spread must rise to attract buyers. That is exactly what happened!!

The figure is a snapshot of the fire sales of assets that occurred due to the panic. Money was lost in these fire sales. To be concrete, suppose the bond was purchased for $100, and then was sold, hoping to fetch $100 (its market value just before the crisis onset). Instead, when all firms are selling the Aaa-rated bonds the price may be, say, $90—a loss of $10. This is how actual losses can occur due to fire sales caused by the panic.

Q. How could this have happened?

A. The development of the parallel banking system did not happen overnight. It has been developing for three decades, and especially grew in the 1990s. But bank regulators and academics were not aware of these developments. Regulators did not measure or understand this development. As we have seen, the government does not measure the relevant markets. Academics were not aware of these markets; they did not study these markets. The incentives of regulators and academics did not lead them to look hard and ask questions.

18.3. SUMMARY

The important points are:

- As traditional banking became unprofitable in the 1980s, due to competition from, most importantly, money market mutual funds and junk bonds, securitization developed. Regulation Q that limited the interest rate on bank deposits was lifted, as well. Bank funding became much more expensive. Banks could no longer afford to hold passive cash flows on their balance sheets. Securitization is an efficient, cheaper, way to fund the traditional banking system. Securitization became sizable.
- The amount of money under management by institutional investors has grown enormously. These investors and non-financial firms have a need for a short-term, safe, interest-earning, transaction account like demand deposits: repo. Repo also grew enormously, and came to use securitization as an important source of collateral.
- Repo is money. It was counted in M3 by the Federal Reserve System, until M3 was discontinued in 2006. But, like other privately-created bank money, it is vulnerable to a shock, which may cause depositors to rationally withdraw en masse, an event which the banking system—in this case the shadow banking system—cannot withstand alone. Forced by the withdrawals to sell assets, bond prices plummeted and firms failed or were bailed out with government money.
- In a bank panic, banks are forced to sell assets, which causes prices to go down, reflecting the large amounts being dumped on the market. Fire sales cause losses. The fundamentals of subprime were not bad enough by themselves to have created trillions in losses globally. The mechanism of the panic triggers the fire sales. As a matter of policy, such firm failures should not be caused by fire sales.
- The crisis was not a one-time, unique, event. The problem is structural. The explanation for the crisis lies in the structure of private transaction securities that are created by banks. This structure, while very important for the economy, is subject to periodic panics if there are shocks that cause concerns about counterparty default. There have been banking panics throughout U.S. history, with private bank notes, with demand deposits, and now with repo. The economy needs banks and banking. But bank liabilities have a vulnerability.

REFERENCES

Appleton, Nathan (1857), Remarks on Currency and Banking: Having Reference to the Present Derangement of the Circulating Medium in the United States (J.H. Eastburn's Press: Boston; reprint of 1841 original).

Gorton, Gary B. (2010), Slapped by the Invisible Hand: The Panic of 2007 (Oxford University Press; 2010).

Gorton, Gary B. (2009a), "Slapped in the Face by the Invisible Hand: Banking and the Panic of 2007," http://papers.ssrn.com/sol3/papers.cfm?abstract id=1401882.

Gorton, Gary B. (2009b), "Information, Liquidity, and the (Ongoing) Panic of 2007," *American Economic Review, Papers and Proceedings*, vol. 99, no. 2 (May 2009), 567–72; http://papers.ssrn.com/sol3/papers.cfm?abstract id=1324195

Gorton, Gary B. and Andrew Metrick (2009a), "Securitized Banking and the Run on Repo," http://papers.ssrn.com/sol3/papers.cfm?abstract id=1440752.

Seligman, Edwin (1908), "The Crisis of 1907 in the Light of History," Introduction to The Currency Problem and the Present Financial Situation, A Series of Addresses Delivered at Columbia University 1907–1908 (Columbia University Press: New York; 1908); p. vii–xxvii.

Vanderlip, Frank (1908), "The Modern Bank," The Currency Problem and the Present Financial Situation, A Series of Addresses Delivered at Columbia University 1907–1908 (Columbia University Press: New York; 1908); p. 1–18.

19

Collateral Crises

GARY B. GORTON AND GUILLERMO ORDOÑEZ*

Financial crises are hard to explain without resorting to large shocks. But the recent crisis, for example, was not the result of a large shock. The Financial Crisis Inquiry Commission (FCIC) Report (2011) noted that with respect to subprime mortgages, "Overall, for 2005 to 2007 vintage tranches of mortgage-backed securities originally rated triple-A, despite the mass downgrades, only about 10 percent of Alt-A and 4 percent of subprime securities had been "materially impaired"—meaning that losses were imminent or had already been suffered—by the end of 2009" (pp. 228–29). Park (2011) calculates the realized principal losses on the $1.9 trillion of AAA/Aaa-rated subprime bonds issued between 2004 and 2007 to be 17 basis points as of February 2011.[1] Though

* We thank Fernando Alvarez, Hal Cole, Tore Ellingsen, Ken French, Mikhail Golosov, Veronica Guerrieri, Todd Keister, Nobu Kiyotaki, David K. Levine, Guido Lorenzoni, Kazuhiko Ohashi, Mario Pascoa, Vincenzo Quadrini, Adriano Rampini, Alp Simsek, Andrei Shleifer, Javier Suarez, Laura Veldkamp, Warren Weber, and seminar participants at Berkeley, Boston College, Columbia GSB, Dartmouth, EIEF, Federal Reserve Board, Maryland, Minneapolis Fed, Ohio State, Princeton, Richmond Fed, Rutgers, Stanford, Wesleyan, Wharton School, Yale, the ASU Conference on "Financial Intermediation and Payments," the Bank of Japan Conference on "Real and Financial Linkage and Monetary Policy," the 2011 SED Meetings at Ghent, the 11th FDIC Annual Bank Research Conference, the Tepper-LAEF Conference on "Advances in Macro-Finance," the Riksbank Conference on "Beliefs and Business Cycles," the 2nd BU/Boston Fed Conference on "Macro-Finance Linkages," The Atlanta Fed Conference on Monetary Economics, the NBER EFG group Meetings in San Francisco, the Banco de Portugal 7th Conference on Monetary Economics, and the 2013 AEA Meetings in San Diego for their comments. We also thank Thomas Bonczek, Paulo Costa, and Lei Xie for research assistance. The authors have nothing to currently disclose, but Gorton was a consultant to AIG Financial Products, 1996–2008.

1. Park (2011) examined the trustee reports from February 2011 for 88.6 percent of the notional amount of AAA subprime bonds issued between 2004 and 2007. The final realized losses on subprime mortgages will not be known for some years. Mortgage securitizations originated in 2006 show the worst losses, but even these are low. Subprime mortgage-backed securities originated in

house prices fell significantly, the effects on mortgage-backed securities, the relevant shock for the financial sector, were not large. But the crisis was large: the FCIC report goes on to quote Ben Bernanke's testimony that of "13 of the most important financial institutions in the United States, 12 were at risk of failure within a period of a week or two" (p. 354). A small shock led to a systemic crisis. The challenge is to explain how a small shock can sometimes have a very large, sudden effect, while at other times the effect of the same size shock is small or nonexistent.

One link between small shocks and large crises is leverage. Financial crises are typically preceded by credit booms, and credit growth is the best predictor of the likelihood of a financial crisis.[2] This suggests that a theory of crises should also explain credit booms. But, since leverage per se is not enough for small shocks to have large effects, it also remains to address what gives leverage its potential to magnify shocks. We develop a theory of financial crises, based on the dynamics of the production and evolution of information in short-term debt markets, that is private money such as (uninsured) demand deposits and money market instruments. As we explain below, we have in mind sale and repurchase agreements (repo) that were at the center of the recent financial crisis. We explain how credit booms arise, leading to financial fragility where a small shock can sometimes have large consequences. In short, "tail risk" is endogenous.

Gorton and Pennacchi (1990) and Dang, Gorton, and Holmström (2013) argue that short-term debt, in the form of bank liabilities or money market instruments, is designed to provide transactions services by allowing trade between agents without fear of adverse selection (due to possible endogenous private information production). In their terminology, this is accomplished by designing debt to be "information-insensitive," that is, such that it is not profitable for any agent to produce private information about the assets backing the debt, the collateral. Adverse selection is avoided in trade. But in a financial crisis there is a sudden loss of confidence in short-term debt in response to a shock. A "loss of confidence" has the precise meaning that the debt becomes information-sensitive; agents may produce information and determine whether the backing collateral is good or not.

We build on these micro foundations to investigate the role of such information-insensitive debt in the macro economy. We do not explicitly model

2006 show realized losses of 1.02 percent through December 2011, and prime MBS originated in 2006 had higher losses, 4.01 percent. See Xie (2012). The "Lehman shock" was endogenous to the crisis; see Gorton, Metrick, and Xie (2012).

2. See, for example, Claessens, Kose, and Terrones (2011), Schularick and Taylor (2012), Reinhart and Rogoff (2009), Borio and Drehmann (2009), Mendoza and Terrones (2008), and Collyns and Senhadji (2002). Jorda, Schularick, and Taylor (2011) (p. 1) study 14 developed countries over 140 years (1870–2008): "Our overall result is that credit growth emerges as the best single predictor of financial instability."

the trading motive for short-term information-insensitive debt. Nor do we explicitly include financial intermediaries. We assume that households have a demand for such debt, and we assume that the short-term debt is issued directly by firms to households to obtain funds and finance efficient projects. Information production about the backing collateral is costly to produce, and agents do not find it optimal to produce (costly) information at every date, which leads to a depreciation of information over time in the economy. We isolate and investigate the macro dynamics of this lack of information production and the possible sudden threat of information production in response to a (possibly small) shock.

The key dynamic in the model concerns how the perceived quality of collateral evolves if (costly) information is not produced. Collateral is subject to idiosyncratic shocks so that over time, without information production, the perceived value of all collateral tends to be the same because of mean reversion toward a "perceived average quality," such that some collateral is known to be bad, but it is not known which specific collateral is bad. Agents endogenously select what to use as collateral. Desirable characteristics of collateral include a high perceived quality and a high cost of information production. In other words, optimal collateral would resemble a complicated, structured claim on housing or land, e.g., a mortgage-backed security.

When information is not produced and the perceived quality of collateral is high enough, firms with good collateral can borrow, but in addition some firms with bad collateral can borrow. In fact, consumption is highest if there is never information production, because then all firms can borrow, regardless of their true collateral quality. The resulting credit boom increases consumption because more and more firms receive financing and produce output. In our setting opacity can dominate transparency, and the economy can enjoy a blissful ignorance. If there has been information-insensitive lending for a long time, that is, information has not been produced for a long time, there is a significant decay of information in the economy—all is gray, there is no black and white—and only a small fraction of true collateral is of known quality.

In this setting we introduce aggregate shocks that may decrease the perceived value of collateral in the economy. Think of the collateral as mortgage-backed securities, for example, being used as collateral for repo, where the households are lending to the firms and receive the collateral. After a credit boom, in which more and more firms borrow with debt backed by collateral of unknown type (but with high perceived quality), a negative aggregate shock affects a larger fraction of collateral than the same aggregate shock would affect when the credit boom was shorter or if the value of collateral was known. Hence, the origin of a crisis is exogenous, but not its size, which depends on how long debt has been information-insensitive in the past and, hence, how large the corresponding boom has been.

A negative aggregate shock may or may not trigger information production. There may be no effect. It depends on the length of the credit boom. If the shock comes after a long enough credit boom, households have an incentive to learn the true quality of the collateral. Then firms may prefer to cut back on the amount borrowed (a credit crunch) to avoid costly information production, a credit constraint. Or, information may be produced, in which case only firms with good collateral can borrow. In either case, output declines when the economy moves from a regime without fear of asymmetric information to a regime where asymmetric information is a real possibility.

In our theory, there is nothing irrational about the credit boom. It is not optimal to produce information every period, and the credit boom increases output and consumption. There is a problem, however, because private agents, using short-term debt, do not care about the future, which is increasingly fragile. A social planner arrives at a different solution because his cost of producing information is effectively lower. For the planner, acquiring information today has benefits tomorrow, which are not taken into account by private agents. When choosing an optimal policy to manage the fragile economy, the planner weighs the costs and benefits of fragility. Fragility is an inherent outcome of using the short-term collateralized debt, and so the planner chooses an optimal level of fragility. This is often popularly discussed in terms of whether the planner should "take the punch bowl away" at the (credit boom) party. Here, the optimal policy may be interpreted as reducing the amount of punch in the bowl, but not taking it away.

Our model is intended to capture the central features of the recent financial crisis. In particular, the crisis was preceded by a credit boom that was ended by a bank run on sale and repurchase agreements (repo) (see Gorton 2010 and Gorton and Metrick 2012a). In a repo transaction a lender lends money at interest, usually overnight, and receives collateral in the form of a bond from the borrower. The collateral is accepted by both parties as recognizably information-insensitive, i.e., no information is produced. Indeed, as in our model much of the collateral was very opaque (i.e., had high information production costs relative to the frequency of the transactions) and was linked to land and housing (subprime bonds). Opacity was the intention of these structures to avoid information production.

In a repo transaction the loan may be overcollateralized; for example, the lender lends $90 but requests collateral with a market value of $100. This is known as a "haircut," 10 percent in this example. If there was no haircut yesterday (a loan for $100 was backed by $100 of collateral), then today there was a withdrawal of $10 from the bank, which must now finance the extra $10 some other way. The financial crisis essentially was this type of bank run; $1.2 trillion was withdrawn in a short period of time (see Gorton and Metrick 2012b). Much of the collateral (we don't know how much) was privately produced securitized

bonds. The subprime shock caused haircuts to rise as lenders questioned the value of the collateral.

Prior to the recent crisis there was a credit boom, particularly in housing. The mortgages were typically securitized into bonds that were used as collateral in repo. During the credit boom, over 1996–2007, nonagency (i.e., private) residential mortgage-backed security issuance grew by 1,248 percent, while commercial mortgage-backed securities grew by 1,691 percent. When house prices started to decline these mortgage-backed securities became questionable, leading to the financial crisis, when the short-term debt was not renewed, leading to almost a complete collapse in the volume of collateral. Over 2007–2012, nonagency residential mortgage-backed securities fell by 100 percent, while commercial mortgage-backed securities fell by 91 percent.[3] The decline in house prices led lenders to question the value of the collateral in mortgage-backed bonds, as well as other securitizations.

We model repo as short-term collateralized debt that firms issue directly to households, abstracting from intermediaries. Indeed, the repo market was not solely an interbank market; see Gorton and Metrick (2012b). As in the financial crisis, nonfinancial firms were dramatically affected as financial intermediaries hoarded cash and refused to lend.[4] In our model we examine this direct impact from the shock to collateral values.

In the model, to rationalize short-term debt and to avoid keeping track of the distribution of land among economic agents, we assume an overlapping generation structure, where agents have a short horizon. Their myopia, however, is the source of a market failure that would not be present in a dynastic structure. The collateral for the short-term debt is called "land" in the model, shorthand for preexisting asset-backed and mortgage-backed securities (MBS). We do not model the primary market or the securitization process. Rather, as time goes by this happens implicitly as new firms offer their land/MBS as collateral. The model displays the dynamics of the crisis, for simplicity, not through higher haircuts but directly through lower credit. There is a lending boom, and then a (small) shock can cause the value of the backing collateral to be questioned.

The crisis corresponds to the case where information is produced and only good collateral can be used once it has been identified. During the financial crisis, some repo collateral was not as affected; it appeared to be "good" collateral. For example, the haircuts on corporate bond collateral were zero (for high-quality dealer banks) before and during the crisis until after the Lehman bankruptcy

3. The source of this information is SIFMA, "US Mortgage-Related Issuance and Outstanding," http://www.sifma.org/research/statistics.aspx.

4. This is documented by, for example, Ivashina and Scharfstein (2010) and Campello, Graham, and Harvey (2010).

when they rose slightly (see Gorton and Metrick 2010). The collateralized loan obligation market was also able to differentiate itself.[5] And, of course, US Treasury bonds continued as collateral during the crisis. In the model a crisis causes output and consumption to drop because there is not enough good collateral to sustain the efficient level of borrowing.

Literature Review—We are certainly not the first to explain crises based on a fragility mechanism. Allen and Gale (2004) define fragility as the degree to which ". . . small shocks have disproportionately large effects." Some literature shows how small shocks may have large effects, and some literature shows how the same shock may sometimes have large effects and sometimes small effects. Our work tackles both aspects of fragility.

Kiyotaki and Moore (1997) show that leverage can have a large amplification effect. This amplification mechanism relies on feedback effects to collateral value over time, while our mechanism is about a sudden informational regime switch. A related literature relies on credit constraints to generate "overborrowing" due to feedback effects from prices on collateral. Leverage increases as the collateral grows in value during an expansion. Then, in some of these settings, private agents do not internalize the effects of their own leverage in depressing collateral prices in the case of shocks that trigger fire sales. Since a shock is an exogenous unlucky event, the policy implications are clear: there should be less borrowing. Examples of this literature include Lorenzoni (2008), Bianchi (2011), and Mendoza (2010).

In contrast to these settings, we explicitly exclude the channel that collateral becomes more valuable due to prices rising, and fire sales are not an issue. In our setting, the effect of the shock occurs only if the credit boom has gone on long enough; the same-sized shock is not always amplified. Furthermore, there is nothing necessarily bad about leverage in our model, and fragility may be the efficient outcome. Other differences are relevant too. First, leverage manifests itself not as more borrowing based on each unit of collateral, but as more units of collateral being able to sustain borrowing. Second, leverage always relaxes endogenous credit constraints. Finally, rather than assuming that a fraction of assets cease to be accepted as collateral, we obtain such a fraction endogenously, microfounding the reduction of credit.

Papers that focus on potential different effects of the same shock are based on equilibrium multiplicity. Diamond and Dybvig (1983), for example, show that banks are vulnerable to random external events (sunspots) when beliefs about the solvency of banks are self-fulfilling.[6] Our work departs from this literature

5. This is a form of securitization where the bonds are backed by bank loans to nonfinancial firms.

6. Other examples include Lagunoff and Schreft (1999), Allen and Gale (2004), and Ordoñez (forthcoming).

because fragility evolves endogenously over time, and it is not based on equilibria multiplicity but on switches between uniquely determined information regimes.

Our article is also related to the literature on leverage cycles developed by Geanakoplos (1996 and 2010) and Geanakoplos and Zame (2010) but highlights the role of information production in fueling those cycles. Furthermore, in our model leverage is not captured by more borrowing from a single unit of collateral, but from more units of collateral in the economy.

There are a number of papers in which agents choose not to produce information ex ante and then may regret this ex post. Examples are the work of Hanson and Sunderam (2013), Pagano and Volpin (2012), Andolfatto (2010), and Andolfatto, Berentsen, and Waller (forthcoming). Like us these models have endogenous information production, but our work describes the endogenous dynamics and real effects of such information.

Two other recent related papers are those of Chari, Shourideh, and Zetlin-Jones (2012) and Guerrieri and Shimer (2012), who discuss adverse selection and asymmetric information as key elements to understanding the recent crisis. In contrast our paper goes one step further and studies the incentives that may induce asymmetric information in the first place.

There is also a recent literature that stresses the role of a rise in firm-level idiosyncratic risk as a contributor of the crisis (e.g., Bigio 2012 and Christiano, Motto, and Rostagno forthcoming). In our model there are two ways to accommodate a mean preserving increase in cross-sectional dispersion. First, an exogenous increase in the dispersion of *perceived* values of collateral, which is an endogenous object in our model, has the same effect of a sudden information acquisition, reducing output. Second, an exogenous increase in the dispersion of *real* values of collateral also reduces output, but its effect is smaller when less information about collateral is available. Even when our model generates a relation between dispersion and output in line with previous work, the effect of perceived values dispersion is endogenous, while the effect of real values dispersion depends on the phase of the credit boom.

In sum, our model produces a "Minsky moment" in which there is an endogenous regime switch causing a crisis, although the mechanism that produces it here is very different from what Minsky had in mind, which was more behavioral (see, e.g., Minsky 1986). From our point of view, a Minsky moment is the idea that emphasizes that a financial crisis is a special event, not just an amplification of a shock. Our mechanism does not rely on a "large" shock.

In the next section we present a single period setting and study the information properties of debt. In Section 19.2 we study the aggregate and dynamic implications of information. We consider policy implications in Section 19.3. In Section 19.4, we conclude.

19.1. A SINGLE PERIOD MODEL

In this section we lay out the basic model in a single period setting. In the next section the model is extended to many periods.

19.1.1. Setting

There are two types of agents in the economy, each with mass 1—firms and households—and two types of goods—*numeraire* and *land*. Agents are risk neutral and derive utility from consuming numeraire at the end of the period. While numeraire is productive and reproducible—it can be used to produce more numeraire—land is not. Since numeraire is also used as *capital* we denote it by K.

Only firms have access to an inelastic fixed supply of nontransferrable managerial skills, which we denote by L^*. These skills can be combined with numeraire in a stochastic Leontief technology to produce more numeraire, K'.

$$K' = \begin{cases} A \min\{K, L^*\} & \text{with prob. } q \\ 0 & \text{with prob. } (1-q). \end{cases}$$

We assume production is efficient, $qA > 1$. Then, the optimal scale of numeraire in production is simply $K^* = L^*$.

Households and firms not only differ in their managerial skills, but also in their initial endowments. On the one hand, households are born with an endowment of numeraire $\bar{K} > K^*$, enough to sustain optimal production in the economy. On the other hand, firms are born with land (one unit of land per firm), but no numeraire.[7]

Even though land is nonproductive, it potentially has an intrinsic value. If land is *"good,"* it delivers C units of numeraire at the end of the period. If land is *"bad,"* it does not deliver any numeraire at the end of the period. We assume a fraction $\hat{p}$ of land is good. At the beginning of the period, the units of land can potentially be heterogeneous in their prior probability of being good. We denote these priors p_i per unit of land i and assume they are common to all agents in the economy. Determining the quality of land with certainty costs γ units of numeraire.

To fix ideas it is useful to think of an example. Assume oil is the intrinsic value of land. Land is good if it has oil underground, which can be exchanged for C units of numeraire at the end of the period. Land is bad if it does not have any oil underground. Oil is nonobservable at first sight, but there is a common perception about the probability each unit of land has oil underground. It is possible to confirm this perception by drilling the land at a cost γ units of numeraire.

7. This is just a normalization. We can alternatively assume firms have an endowment of numeraire $\bar{K}_{firms}$, but not enough to finance optimal production $\bar{K}_{firms} < K^* < \bar{K} + \bar{K}_{firms}$.

In this simple setting, resources are in the wrong hands. Households have only numeraire while firms have only managerial skills, but production requires that both inputs be in the same hands. Since production is efficient, if output were verifiable it would be possible for firms to borrow the optimal amount of numeraire K^* by issuing state contingent claims. In contrast, if output were nonverifiable, firms would never repay, and households would never be willing to lend.

We focus on this latter case in which firms can hide numeraire but cannot hide land, which renders land useful as *collateral*. Firms can commit to transfer a fraction of land to households if they do not repay the promised numeraire, which relaxes the financial constraint imposed by the nonverifiability of output.

The perception about the quality of collateral then becomes critical in facilitating credit. We assume that $C > K^*$, which implies that land that is known to be good can sustain the optimal loan size, K^*. In contrast, land that is known to be bad cannot sustain any loan.[8] But how much can a firm with a piece of land that is good with probability p borrow? Is information about the true value of land produced or not?

19.1.2. Optimal Loan for a Single Firm

In this section we study the optimal short-term collateralized debt for a single firm, considering the possibility that households may want to produce information about the land posted as collateral. In this article we study a single-sided information problem, since the firm does not have resources in terms of numeraire to learn about the collateral. In a companion paper, Gorton and Ordoñez (2013) extend the model to allow both borrowers and lenders to be able to acquire information about collateral.

We impose two assumptions. First, lenders' acquisition of information and the information itself become public only at the end of the period, unless lenders decide to disclose it earlier. This implies that asymmetric information can potentially exist during the period. Second, each firm is randomly matched with a household and the firm has the negotiation power in determining the loan conditions. In the Appendix we show that explicitly modeling competition across lenders complicates the exposition and only strengthens our results.

8. Since we assume $C > K^*$, the issue arises of whether a firm with an excess of good collateral can sell land to another firm with bad collateral to finance optimal borrowing in the economy. We rule this out, implicitly assuming that the firm with good land has to hold the whole unit of land to maintain its value, which renders collateral ownership effectively indivisible. Empirically, for example, if the originator, sponsor, and servicer of a mortgage-backed security is the same firm, the collateral has a higher value compared to the situation in which these roles are separated in different firms. See Demiroglu and James (2012).

Firms optimally choose between debt that triggers information acquisition about the collateral (*information-sensitive debt*) or not (*information-insensitive debt*). Triggering information acquisition is costly because it raises the cost of borrowing to compensate for the monitoring cost γ. However, not triggering information acquisition may also be costly because it may imply less borrowing to discourage households from producing information. This trade-off determines the information-sensitiveness of the debt and, ultimately, the volume and dynamics of information in the economy.

19.1.2.1 Information-Sensitive Debt

Under this contract, lenders learn the true value of the borrower's land by paying an amount γ of numeraire, and loan conditions are conditional on the resulting information. Since by assumption lenders are risk neutral and break even,

$$p\left(qR_{IS} + \left(1-q\right)x_{IS}C - K\right) = \gamma\,, \tag{19.1}$$

where K is the size of the loan, R_{IS} is the face value of the debt, and x_{IS} is the fraction of land posted by the firm as collateral.

The firm should pay the same in case of success or failure. If $R_{IS} > x_{IS}C$, the firm would always default, handing over the collateral rather than repaying the debt. In contrast, if $R_{IS} < x_{IS}C$ the firm would always sell the collateral directly at a price C and repay lenders R_{IS}. In this setting, then, debt is risk free, which renders the results under risk neutrality to hold without loss of generality. This condition pins down the fraction of collateral that a firm posts as a function of p,

$$R_{IS} = x_{IS}C \Rightarrow x_{IS} = \frac{pK+\gamma}{pC} \leq 1.$$

It is feasible for firms to borrow the optimal scale K^* only if $\frac{pK^*+\gamma}{pC} \leq 1$, or if $p \geq \frac{\gamma}{C-K^*}$. If this is not the case, firms can borrow only $K = \frac{pC-\gamma}{p} < K^*$ when posting the whole unit of good land as collateral. Finally, it is not feasible to borrow at all if $pC < \gamma$.

Expected profits net of the land value pC from information-sensitive debt are

$$E\left(\pi \,|p, IS\right) = p\left(qAK - x_{IS}C\right),$$

and using x_{IS} from above,

$$E\left(\pi \,|p, IS\right) = pK^*\left(qA-1\right) - \gamma\,. \tag{19.2}$$

Intuitively, with probability p collateral is good and sustains expected production of $K^*\left(qA-1\right)$ of numeraire, and with probability $\left(1-p\right)$ collateral is bad and does not sustain any loan or production. However, the firm always has to compensate in expectation for the monitoring costs, γ.

It is profitable for firms to borrow the optimal scale inducing information as long as $pK^*\left(qA-1\right) \geq \gamma$, or $p \geq \frac{\gamma}{K^*\left(qA-1\right)}$. Combining the profitability and feasibility conditions, if $\frac{\gamma}{K^*\left(qA-1\right)} > \frac{\gamma}{C-K^*}$ $\left(\text{or } qA < C/K^*\right)$, whenever the firm wants to borrow, it is feasible to borrow the optimal scale K^* if the land is found to be good. Simply to minimize the kinks in the firm's profit function, we assume this condition holds

$$E\left(\pi \left| p, IS\right.\right) = \begin{cases} pK^*\left(qA - 1\right) - \gamma & \text{if } p \geq \frac{\gamma}{K^*\left(qA-1\right)} \\ 0 & \text{if } p < \frac{\gamma}{K^*\left(qA-1\right)}. \end{cases}$$

19.1.2.2 Information-Insensitive Debt

Another possibility is for firms to borrow without triggering information acquisition. Again, since by assumption lenders are risk neutral and break even,

$$qR_{II} + \left(1 - q\right) px_{II}C = K, \tag{19.3}$$

subject to debt being risk free, $R_{II} = x_{II}pC$ for the same reasons as above. Then

$$x_{II} = \frac{K}{pC} \leq 1.$$

For this contract to be information-insensitive, borrowers should be confident that lenders do not have incentives to deviate, secretly checking the value of collateral and lending only if the collateral is good, pretending that they do not know the collateral value. Lenders do not want to deviate if the expected gains from acquiring information, evaluated at x_{II} and R_{II}, are less than its costs, γ. Formally,

$$p\left(qR_{II} + \left(1 - q\right)x_{II}C - K\right) < \gamma \Rightarrow \left(1-p\right)\left(1 - q\right)K < \gamma.$$

Intuitively, by acquiring information the lender lends only if the collateral is good, which happens with probability p. If there is default, which occurs with probability $\left(1-q\right)$, the lender can sell at $x_{II}C$ of collateral that was effectively purchased at $K = px_{II}C$, making a net gain of $\left(1-p\right)x_{II}C = \left(1-p\right)\frac{K}{p}$.

It is clear from the previous condition that the firm can discourage information acquisition by reducing borrowing. If the condition does not bind when evaluated at $K = K^*$, there are no incentives for lenders to produce information. In contrast, if the condition binds, the firm will borrow as much as possible given the restriction of not triggering information acquisition:

$$K = \frac{\gamma}{\left(1-p\right)\left(1 - q\right)}. \tag{19.4}$$

Even though the technology is linear, the constraint on borrowing has p in the denominator, which induces convexity in expected profits.

Information-insensitive borrowing is characterized by the following debt size:

$$K\left(p\,|II\right) = \min\left\{K^*, \frac{\gamma}{\left(1-p\right)\left(1-q\right)}, pC\right\}. \tag{19.5}$$

That is, borrowing is either constrained technologically (there are no credit constraints, but firms do not need to borrow more than K^*), informationally (there are credit constraints and firms cannot borrow more than $\frac{\gamma}{(1-p)(1-q)}$ without triggering information production) or by low collateral value (the unit of land is not worth more than pC).

Expected profits net of the land value pC for information-insensitive debt are

$$E\left(\pi\,\middle|p,II\right) = qAK - x_{II}pC,$$

and using x_{II}

$$E\left(\pi\,\middle|p,II\right) = K\left(p\,|II\right)\left(qA-1\right). \tag{19.6}$$

Considering the kinks explicitly, these profits are

$$E\left(\pi\,\middle|p,II\right) = \begin{cases} K^*\left(qA-1\right) & \text{if } K^* \leq \frac{\gamma}{(1-p)(1-q)} \text{ (no credit constraint)} \\ \frac{\gamma}{(1-p)(1-q)}\left(qA-1\right) & \text{if } K^* > \frac{\gamma}{(1-p)(1-q)} \text{ (credit constraint)} \\ pC(qA-1) & \text{if } pC < \frac{\gamma}{(1-p)(1-q)} \text{ (low collateral value)}. \end{cases}$$

The first kink is generated by the point at which the constraint to avoid information production is binding when evaluated at the optimal loan size K^*; this occurs when financial constraints start binding more than technological constraints. The second kink is generated by the constraint $x_{II} \leq 1$, under which the firm is not constrained by the threat of information acquisition, but it is directly constrained by the low expected value of the collateral, pC.

19.1.2.3 Induce Information Acquisition or Not?

Depending on the belief p about its collateral, a firm compares equations (19.2) and (19.6) to choose between issuing information-insensitive debt (II) or information-sensitive debt (IS). The proof of the next proposition is trivial. The proofs of all other propositions are in the Appendix.

PROPOSITION 1: *Firms borrow inducing information acquisition if*

$$\frac{\gamma}{qA-1} < pK^* - K\left(p\,|II\right), \tag{19.7}$$

and without inducing information acquisition otherwise.

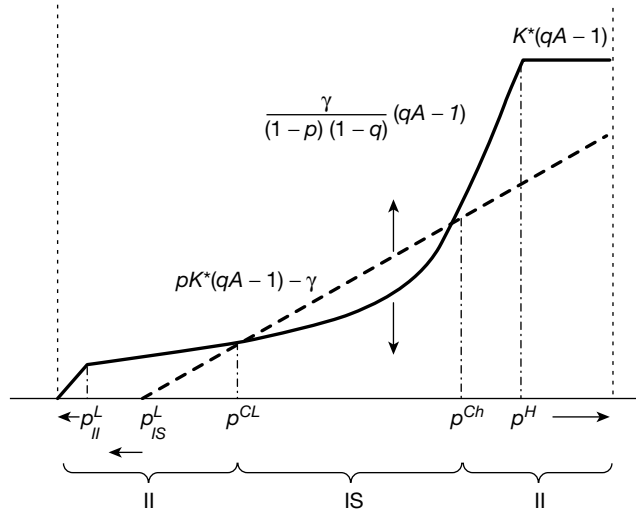


Figure 19.1 Single Period Expected Profits

Figure 19.1 shows the ex ante expected profits, net of the expected value of land, under the two information regimes, for each possible p.

The cut offs highlighted in Figure 19.1 are determined in the following way: The cut off p^H is the belief that generates the first kink of information-insensitive profits, below which firms have to reduce borrowing to prevent information acquisition:

$$p^H = 1 - \frac{\gamma}{K^*\left(1-q\right)}. \tag{19.8}$$

The cut off p^L_{II} comes from the second kink of information-sensitive profits:[9]

$$p^L_{II} = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{\gamma}{C\left(1-q\right)}}. \tag{19.9}$$

The cut off p^L_{IS} comes from the kink of information-sensitive profits:

$$p^L_{IS} = \frac{\gamma}{K^*\left(qA - 1\right)}. \tag{19.10}$$

Cutoffs p^{Ch} and p^{Cl} are obtained from equalizing the profit functions under information-sensitive and -insensitive debt, and solving the quadratic equation:

9. The positive root for the solution of $pC = \gamma / \left(1-p\right)\left(1-q\right)$ is irrelevant since it is greater than p^H, and then firms are not credit but technologically constrained, just borrowing K^*.

$$\gamma = \left[pK^* - \frac{\gamma}{(1-p)(1-q)} \right] (qA - 1). \qquad (19.11)$$

Information-insensitive loans are chosen for collateral with high and low beliefs p. Information-sensitive loans are chosen for collateral with intermediate values of p. The first regime generates symmetric ignorance about the value of collateral. The second regime generates symmetric information about the value of collateral.

How do these regions depend on information costs? The five arrows in Figure 19.1 show how the cut offs and functions move as we reduce γ. If information is free ($\gamma = 0$), all collateral is information-sensitive (i.e., the IS region is $p \in [0,1]$). As γ increases, the two cut offs p^{Ch} and p^{Cl} converge, and the IS region shrinks until it disappears when γ is large enough (i.e., the II region is $p \in [0,1]$ when $\gamma > \frac{K^*}{C}(C - K^*)$).

Then, conditional on γ, the feasible borrowing for each belief p follows the schedule

$$K(p) = \begin{cases} K^* & \text{if } p^H < p \\ \frac{\gamma}{(1-p)(1-q)} & \text{if } p^{Ch} < p < p^H \\ pK^* - \frac{\gamma}{(qA-1)} & \text{if } p^{Cl} < p < p^{Ch} \\ \frac{\gamma}{(1-p)(1-q)} & \text{if } p^L_{II} < p < p^{Cl} \\ pC & \text{if } p < p^L_{II}. \end{cases} \qquad (19.12)$$

19.1.3. The Choice of Collateral

In this section, in addition to heterogeneous beliefs, p, about land value, we assume land is also heterogenous in terms of the cost γ of acquiring information. What is the combination of p and γ that allows for the largest loans? The next proposition summarizes the answer.

PROPOSITION 2: *Effects of p and γ on borrowing.*

Consider collateral characterized by the pair (p, γ). The reaction of borrowers to these variables depends on financial constraints and information sensitiveness.

(*i*) *Fix γ.*

(*a*) *No financial constraint: Borrowing is independent of p;*
(*b*) *Information-sensitive regime: Borrowing is increasing in p;*
(*c*) *Information-insensitive regime: Borrowing is increasing in p.*

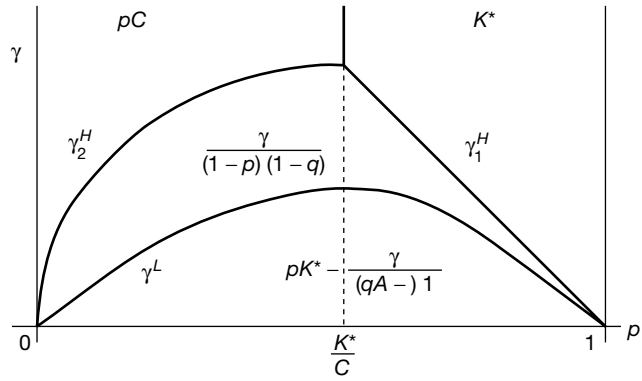


Figure 19.2 Borrowing for Different Types of Collateral

(*ii*) *Fix p.*

(*a*) *No financial constraint: Borrowing is independent of* γ ;
(*b*) *Information-sensitive regime: Borrowing is decreasing in* γ ;
(*c*) *Information-insensitive regime: Borrowing is increasing in* γ *if higher than pC and independent of* γ *if pC.*

Figure 19.2 shows the borrowing possibilities for all combinations (p, γ) and the regions described in Proposition 2 (K^* is the loan without financial constraints, $pK^* - \frac{\gamma}{(qA-1)}$ is the loan in the *IS* regime, while $\frac{\gamma}{(1-p)(1-q)}$ and pC are the loans in the *II* regime).

If it were possible for borrowers to choose the lenders' difficulty in monitoring collateral with belief p, then they would set $\gamma > \gamma_1^H(p)$ for that p, such that $p > p^H(\gamma)$ and the borrowing is K^*, without information acquisition.

This analysis suggests that, endogenously, an economy would be biased towards using collateral with relatively high p and relatively high γ. Agents in an economy will first use collateral that is perceived to be of high quality. As the needs for collateral increase, agents start relying on collateral of worse and worse quality. To accommodate this collateral of poorer expected quality, agents may need to increase γ, making information acquisition difficult and expensive. While outside the scope of our article, this framework can shed light on security design and the complexity of modern financial instruments.

19.1.4. Aggregation

Consider a match between a household and a firm with land that is good with probability p. The expected consumption of a household is $\bar{K} - K(p) + E(repay \mid p)$, and the expected consumption of a firm is $E(K' \mid p) - E(repay \mid p)$.

Aggregate consumption is the sum of the consumption of all households and firms. Since $E\left(K' \middle| p\right) = qAK\left(p\right)$,

$$W_t = \bar{K} + \int_0^1 K\left(p\right)\left(qA - 1\right) f\left(p\right) dp,$$

where $f\left(p\right)$ is the distribution of beliefs about collateral types in the economy and $K\left(p\right)$ is monotonically increasing in p (equation 19.12).

In the unconstrained first-best (the case of verifiable output, for example) all firms borrow K^* and operate at the optimal scale, regardless of beliefs p about the collateral. This implies that the unconstrained first-best aggregate consumption is

$$W^* = \bar{K} + K^*\left(qA - 1\right).$$

Since collateral with relatively low p is not able to sustain loans of K^*, the deviation of consumption from the unconstrained first-best critically depends on the distribution of beliefs p in the economy. When this distribution is biased toward low perceptions about collateral values, financial constraints hinder total production. The distribution of beliefs introduces heterogeneity in production, purely given by heterogeneity in collateral and financial constraints, not by heterogeneity in technological possibilities.

In the next section we study how this distribution of p endogenously evolves over time, and how that affects the dynamics of aggregate production and consumption.

19.2. DYNAMICS

In this section we nest the previous analysis for a single period in an overlapping generations economy. The purpose is to study the evolution of the distribution of collateral beliefs that determines the level of production in the economy in each period.

We assume that each unit of land changes quality over time, mean reverting toward the average quality of land in the economy, and we study how endogenous information acquisition shapes the distribution of beliefs over time. First, we study the case without aggregate shocks to land, in which the average quality of collateral in the economy does not change, and discuss the effects of endogenous information production on the dynamics of credit. Then, we introduce aggregate shocks that reduce the average quality of land in the economy and study the effects of endogenous information acquisition on the size of crises and the speed of recoveries.

19.2.1. Extended Setting

We assume an overlapping generations structure. Every period is populated by two cohorts of individuals who are risk neutral and live for two periods. These individuals are born as households (when *"young"*), with a numeraire endowment of $\bar{K}$ but no managerial skills, and then become firms when *"old,"* with managerial skills L^*,. but no numeraire to use in production. We assume the numeraire is nonstorable and land is storable until the moment its intrinsic value (either C or 0) is extracted, after which the land disappears. This implies that as long as land is transferred, its potential value as collateral remains. As in the single period model, we still assume there is random matching between a firm and a household in every period. The timing is as follows:

- At the beginning of the period land that is good with probability p_{-1} may suffer idiosyncratic or aggregate shocks that move this probability to p.
- After the shocks, each member of the "young" generation (households) matches with a member of the "old" generation (firms) with land that is good with probability p. The household determines the conditions of a loan (pairs $(R_{II}; x_{II})$ and $(R_{IS}; x_{IS})$) that make him indifferent between lending or not (conditions 1 and 3). The firm then chooses a lending contract that maximizes profits selecting the maximum between $E(\pi \mid p, IS)$ and $E(\pi \mid p, II)$ (equations 2 and 6) and begins production. Depending on whether there is information acquisition or not beliefs are updated to zero (bad land) or one (good land) or remain at p, respectively.
- At the end of the period, the firm can choose to sell its unit of land (or the remaining land after default) to the household at a price $Q(p)$ or to extract and consume its intrinsic value.

The optimal loan contract follows the characterization described in the single period model above. The market for land is new. Land can be transferred across generations, and agents want to buy land when young to use it as collateral to borrow productive numeraire when old. This is reminiscent of the role of fiat money in overlapping generations, with the critical differences that land is intrinsically valuable and is subject to imperfect information about its quality. Still, as in those models, we have multiple equilibria based on multiple paths of rational expectations about land prices that incorporate the use of land as collateral.

However, in this article we are not interested in credit booms, bubbles or crises arising from transitions across multiple equilibria, which are typical features of those models. So, we impose restrictions to select the equilibrium in

which the land price just reflects the expected intrinsic value of land when it can be used as collateral (that is, the price of a unit of land with belief p is just $Q(p) = pC$). Choosing this particular equilibrium has the advantage of isolating the dynamics generated by information acquisition.[10]

The first restriction is that information can be produced only at the beginning of the period, not at the end. This assumption means that firms prefer to post land as collateral rather than sell land with the risk of information production. The second restriction is that buyers (households) make take-it-or-leave-it offers for the land of their matched firm at the end of the period; households have all the bargaining power. This implies that sellers will be indifferent between selling the unit of land at pC or consuming pC in expectation. As we discuss in the Appendix, we can characterize the competitive environment to sustain this assumption.

Under these assumptions, the single-period analysis from the previous section just repeats over time. The only changing state variable linking periods is the distribution of beliefs about collateral. We can now define the equilibrium.

DEFINITION 1 (Definition of Equilibrium):

In each period, for each match of a household and a firm of type p an equilibrium is:

- *A pair of debt face values* (R_{II} and R_{IS}) *and a pair of fractions of land to be collected in case of default* (x_{II} and x_{IS}) *such that lenders are indifferent; and a profit maximizing choice of information-sensitive debt or information-insensitive debt.*
- *A land price* $Q(p)$ *is determined by take-it-or-leave-it offer by the household.*
- *Beliefs are updated after information or shocks, using Bayes' rule.*

Next we study the interaction between shocks to collateral and information acquisition to study the dynamics of production in the economy. First we imposed a simple mean reverting process of idiosyncratic shocks and show that information may vanish over time, generating a credit boom sustained by increased symmetric ignorance in the economy. Then, we allow for an unexpected aggregate shock that may introduce the threat of information acquisition and generate crises.

10. Still, our results are robust since the information dynamics that we focus on remain an important force in the other equilibria we ruled out, as long as the price of land increases with p. In the Appendix, we discuss the multiplicity of land prices.

This is the main advantage of focusing on the equilibrium in which the price of collateral just reflects its intrinsic value, and not the future value of collateral. First, credit booms do not arise from bubbles in the price of each unit of collateral, but from an increase in the volume of land that can be used as collateral. Second, credit crises are not generated by shifting from a good to a bad equilibrium, but by shifting from the information-insensitive to the information-sensitive regime that coexist in a unique equilibrium.

19.2.2. No Aggregate Shocks

Here we just introduce idiosyncratic shocks to collateral. We impose a specific process of idiosyncratic mean reverting shocks that are useful in characterizing analytically the dynamic effects of information production on aggregate consumption. First, we assume that the idiosyncratic shocks are observable, but their realization is not observable, unless information is produced. Second, we assume that the probability that a unit of land faces an idiosyncratic shock is independent of land type. Finally, we assume that the probability a unit of land becomes good, conditional on having an idiosyncratic shock, is also independent of its type. These assumptions just simplify the exposition, and the main results are robust to different processes, as long as there is mean reversion of collateral in the economy.

Formally, in each period either the true quality of each unit of land remains unchanged with probability λ, or there is an idiosyncratic shock that changes its type with probability $(1-\lambda)$. In this last case, land becomes good with a probability $\hat{p}$, independent of its current type. Even when the shock is observable, its realization is not, unless a certain amount of the numeraire good γ is used to learn about it.[11].

In this simple stochastic process for idiosyncratic shocks, and in the absence of aggregate shocks to $\hat{p}$, this distribution has a three-point support: 0, $\hat{p}$, and 1. The next proposition shows that the evolution of aggregate consumption depends on $\hat{p}$, which can be either in the information-sensitive or in the information-insensitive region.

PROPOSITION 3 (Evolution of Aggregate Consumption in the Absence of Aggregate Shocks): *Assume there is perfect information about land types in the initial period. If $\hat{p}$ is in the information-sensitive region ($\hat{p} \in [p^{Cl}, p^{Ch}]$), consumption is constant over time and is lower than the unconstrained first-best. If*

11. To guarantee that all land is traded, households should have enough resources to buy good land, $\bar{K} > C$, and they should be willing to pay C for good land even when facing the probability that it may become bad next period, with probability $(1-\lambda)$. Since this fear is the strongest for good land, the sufficient condition is enough persistence of collateral, $\lambda\left(K^*\left(qA-1\right)+C\right) > C$.

$\hat{p}$ is in the information-insensitive region, consumption grows over time if $\hat{p} > \hat{p}_h^*$ or $\hat{p} < \hat{p}_l^*$ where $\hat{p}_l^*$ and $\hat{p}_h^*$ are the solutions to the quadratic equation $\hat{p}^* K^* = \frac{\gamma}{(1-\hat{p}^*)(1-q)}$.

This result is particularly important if the economy has collateral such that $\hat{p} > p^H > \hat{p}_h^*$. In this case consumption grows over time toward the unconstrained first-best. When $\hat{p}$ is high enough, the economy has enough good collateral to sustain production at the optimal scale. As information vanishes over time good collateral implicitly subsidizes bad collateral, and after enough periods virtually all firms are able to produce at the optimal scale, not just those firms with good collateral.

19.2.3. Aggregate Shocks

Now we introduce negative aggregate shocks that transform a fraction $(1 - \eta)$ of good collateral into bad collateral. As with idiosyncratic shocks, the aggregate shock is observable, but which good collateral changes type is not. When the shock hits, there is a downward revision of beliefs about all collateral. That is, after the shock, collateral with belief $p = 1$ gets revised downwards to $p' = \eta$, and collateral with belief $p = \hat{p}$ gets revised downwards to $p' = \eta\hat{p}$.

Based on the discussion about the endogenous choice of collateral, which justifies that collateral would be constructed to maximize borrowing and prevent information acquisition, we focus on the case where, prior to the negative aggregate shock, the average quality of collateral is good enough such that there are no financial constraints (that is, $\hat{p} > p^H$).

In the next proposition we show that the longer the economy does not face a negative aggregate shock, the larger the consumption loss when such a shock does occur.

PROPOSITION 4 (The Larger the Credit Boom and the Shock, the Larger the Crisis): *Assume $\hat{p} > p^H$, and a negative aggregate shock η hits after t periods of no aggregate shocks. The reduction in consumption $\Delta(t|\eta) \equiv W_t - W_{t|\eta}$ is nondecreasing in the size of the shock η and nondecreasing in the time t elapsed previously without a shock.*

The intuition for this proposition is the following. Pooling implies that bad collateral is confused with good collateral. This allows for a credit boom because firms with bad collateral get credit that they would not otherwise obtain. Firms with good collateral effectively subsidize firms with bad collateral since good collateral still gets the optimal leverage, while bad collateral is able to leverage more.

However, pooling also implies that good collateral is confused with bad collateral. This puts good collateral in a weaker position in the event of negative

aggregate shocks. Without pooling, a negative shock reduces the belief that collateral is good from $p = 1$ to $p' = \eta$. With pooling, a negative shock reduces the belief that collateral is good from $p = \hat{p}$ to $p' = \eta\hat{p}$. Good collateral gets the same credit regardless of having beliefs $p = 1$ or $p = \hat{p}$. However, credit may be very different when $p = \eta$ and $p = \eta\hat{p}$. In particular, after a negative shock to collateral, credit may decline since either a high amount of the numeraire needs to be used to produce information, or borrowing needs to be excessively constrained to avoid such information production.

If we define *"fragility"* as the probability that aggregate consumption declines more than a certain value, then the next corollary immediately follows from Proposition 4.

COROLLARY 1: *Given a negative aggregate shock, the fragility of an economy increases with the number of periods the debt in the economy has been informationally insensitive, and, hence, increases with the fraction of collateral that is of unknown quality.*

Proposition 3 describes how information deterioration may induce credit booms, and Proposition 4 describes how the threat of information acquisition may induce crises. What happens next? How does information production affect the speed of recovery?

PROPOSITION 5 (Information and Recoveries): *Assume $\hat{p} > p^H$ and that a negative aggregate shock η generates a crisis in period t. The recovery from the crisis is faster if information is generated after the shock when $\eta\hat{p} < \overline{\eta\hat{p}} \equiv \frac{1}{2} + \sqrt{\frac{1}{4} - \frac{\gamma}{K^*(1-q)}}$, where $p^{Ch} < \overline{\eta\hat{p}} < p^H$. That is, $W^{IS}_{t+1} > W^{II}_{t+1}$* for all *$\eta\hat{p} < \overline{\eta\hat{p}}$* and *$W^{IS}_{t+1} \leq W^{II}_{t+1}$ otherwise.*

The intuition for this proposition is the following. When information is acquired after a negative shock, not only are a lot of resources being spent in acquiring information but also only a fraction $\eta\hat{p}$ of collateral can sustain the maximum borrowing K^*. When information is not acquired after a negative shock, collateral that remains with belief $\eta\hat{p}$ will restrict credit in the following periods, until mean reversion moves beliefs back to $\hat{p}$. This is equivalent to restricting credit proportional to monitoring costs in subsequent periods. Not producing information causes a kind of "lack of information overhang" going forward. The proposition generates the following Corollary.

COROLLARY 2: *There exists a range of negative aggregate shocks (η such that $\eta\hat{p} \in [p^{Ch}, \overline{\eta\hat{p}}]$) in which agents do not acquire information, but recovery would be faster if they did.*

Finally, the next proposition describes the evolution of the standard deviation of beliefs in the economy during credit booms and credit crises.

PROPOSITION 6 (Dispersion of Beliefs During Booms and Crises): *During a credit boom, the standard deviation of beliefs declines. During a credit crisis, if the aggregate shock η triggers information production about collateral with belief $\eta\hat{p}$, the standard deviation of beliefs increases. This increase is larger the longer was the preceding boom.*

Intuitively, credit booms are generated by vanishing information. Since over that process beliefs accumulate to the average quality $\hat{p}$, the dispersion of the belief distribution declines. If this process developed long enough, an aggregate shock that triggers information reveals the true type of most land, and beliefs return to $p = 0$ and $p = 1$ increasing the dispersion of the belief distribution. This effect is stronger the longer the preceding boom that accumulated collateral with beliefs $\hat{p}$.

19.2.4. Numerical Illustration

Now we illustrate our dynamic results with a numerical example. We assume idiosyncratic shocks happen with probability $(1 - \lambda) = 0.1$, in which case the collateral becomes good with probability $\hat{p} = 0.92$. Other parameters are $q = 0.6$, $A = 3$ (investment is efficient and generates a return of 80 percent in expectation), $\bar{K} = 20$, $L^* = K^* = 7$, $C = 15$ (the endowment is large enough to provide a loan for the optimal scale of production and to buy the most expensive unit of land), and $\gamma = 0.35$ (information costs are 5 percent of the optimal loan).

Given these parameters we can obtain the relevant cut offs for our analysis. Specifically, $p^H = 0.88$, $p_{II}^L = 0.06$, and the information-sensitive region of beliefs is $p \in [0.22, 0.84]$. Figure 19.3 plots the ex ante expected profits with information-sensitive (dotted) and -insensitive (solid) debt, and the respective cut offs.

Using these cut offs in each period, we simulate the model for 100 periods. At period zero we assume perfect information about the true quality of each unit of land in the economy. Unless replenished, information vanishes over time due to idiosyncratic shocks. The dynamics of production mirror those of the belief distribution.

In periods 5 and 50 we perturb the economy by introducing negative aggregate shocks that transform a fraction $(1 - \eta)$ of good collateral into bad collateral. We consider shocks of different size, ($\eta = 0.97$, $\eta = 0.91$, and $\eta = 0.90$) and compute the dynamic reaction of aggregate production to them. We choose the size of these shocks to guarantee that $\eta\hat{p}$ is above p^H when $\eta = 0.97$, is between p^{Ch} and p^H when $\eta = 0.91$, and is less than p^{Ch} when $\eta = 0.90$.

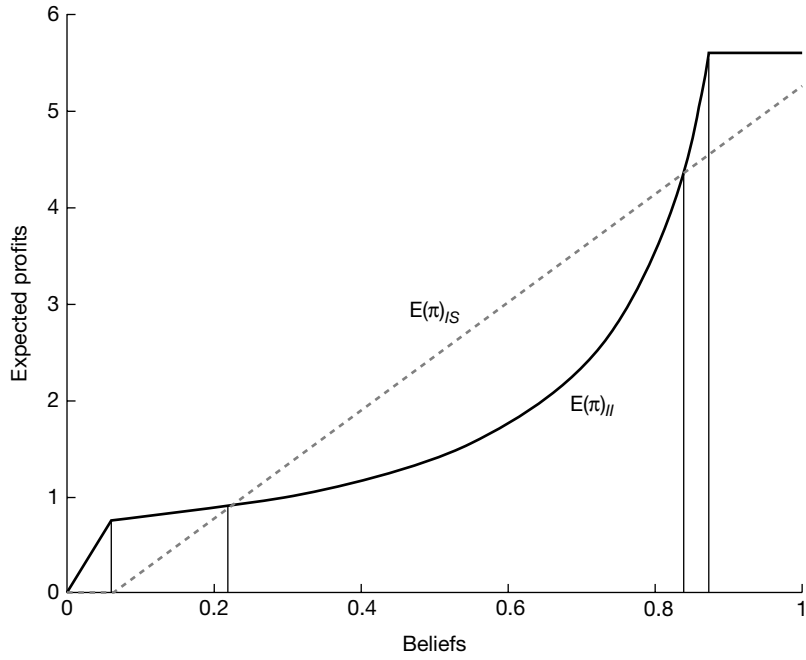


Figure 19.3 Expected Profits and Cut offs

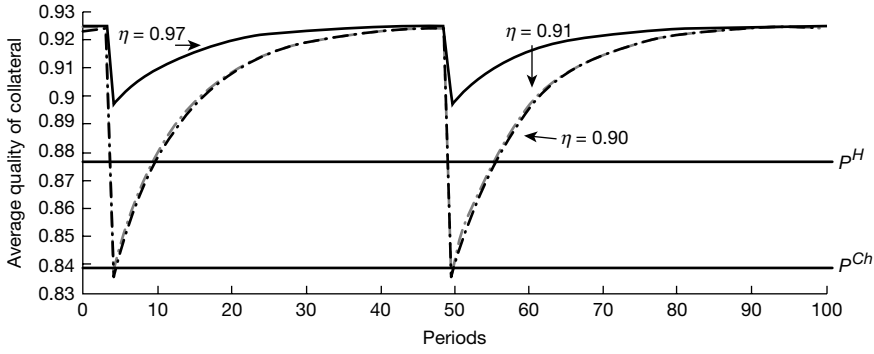


Figure 19.4 Average Quality of Collateral

Figure 19.4 shows the evolution of the average quality of collateral for the three negative aggregate shocks. Since mean reversion guarantees that average quality converges back to $\hat{p} = 0.92$ after the shocks, their effects are only temporary.

Figure 19.5 shows the evolution of aggregate production for the three negative aggregate shocks. A couple of features are worth noting. First, if $\eta = 0.97$, the aggregate shock is so small that it never constrains borrowing or modifies the evolution of production. Second, as proved in Proposition 4, if $\eta = 0.91$ or $\eta = 0.90$, aggregate production drops more in period 50, when the credit boom is

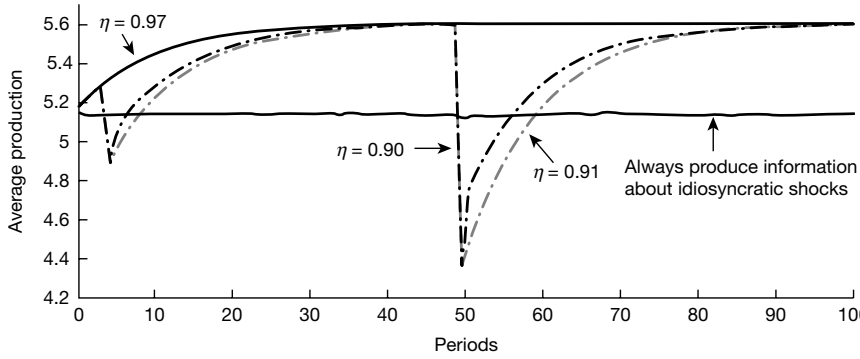


Figure 19.5 Aggregate Production

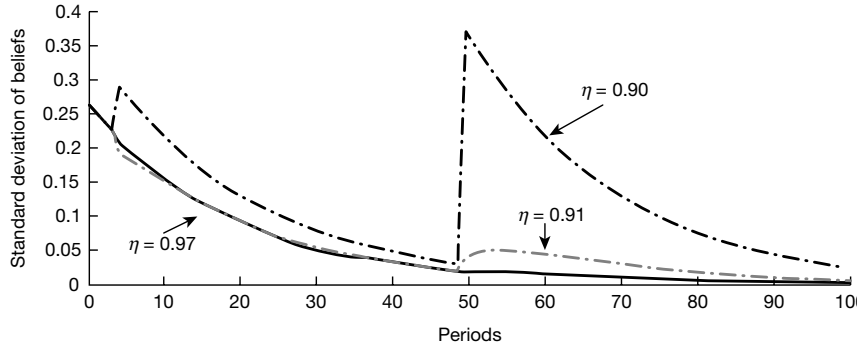


Figure 19.6 Standard Deviation of Distribution of Beliefs

mature and information is scarce, than in period 5, when there is still a large volume of information about collateral in the economy. Critically, the crisis is larger in period 50, not only because it finishes a large boom, but also because credit drops to a lower level. Indeed, aggregate production in period 50 is lower than in period 5 because credit dries up for a larger fraction of collateral when information is scarcer.

As proved in Proposition 5, a shock $\eta = 0.91$ does not trigger information production, but a shock $\eta = 0.90$ does. Even when these two shocks generate production drops of similar magnitude, recovery is faster when the shock is slightly larger and information is replenished.

Figure 19.6 shows the evolution of the beliefs' dispersion, a measure of information availability. As proved in Proposition 6, a credit boom is correlated with a decline in the dispersion of beliefs and, given that after many periods without a shock most collateral looks the same, the information acquisition triggered by a shock $\eta = 0.90$ generates a larger increase in dispersion in period 50.

Finally, to illustrate the negative side of information, Figure 19.7 shows the evolution of production under two very extreme cases: information acquisition

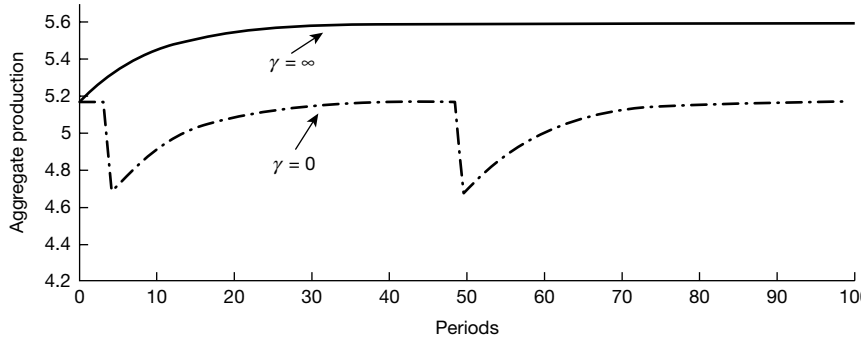


Figure 19.7 Extreme Information Costs

is free ($\gamma = 0$), and it is impossible ($\gamma = \infty$). Aggregate production is lower and more volatile when information is free. It is lower because only firms with good collateral get loans. It is more volatile because the volume of good collateral is subject to aggregate shocks. When information acquisition is free, the reaction of credit is independent of the length of the preceding boom and depends only on the size of the shock. In contrast, when information acquisition is impossible, over time all land is used as collateral, and shocks do not introduce any fear that someone will acquire information and lead to a credit decline.

19.3. POLICY IMPLICATIONS

In this section we discuss optimal information production when a planner cares about the discounted consumption of all generations and faces the same information restrictions and costs as households and firms. More specifically, welfare is measured by

$$U_t = E_t \sum_{\tau=t}^{\infty} \beta^{\tau - t} W_t. \tag{19.13}$$

The planner chooses an endowment transfer (loan size) from households to firms and decides whether or not to generate information about firms' collateral, facing two types of constraints. First, *collateral constraints* prevent the planner from lending a firm more endowment than the expected value of the firm's collateral. This is

$$K(p) \leq \min\left\{K^*, pC\right\}. \tag{19.14}$$

Second, *information constraints* prevent the planner from lending to a firm without acquiring information, if the loan would have triggered information acquisition by private agents in a decentralized economy. This implies that the planner

cannot lend a firm more than the amount in equation (19.4) without acquiring costly information. Then, if

$$K(p) > \frac{\gamma}{(1-p)(1-q)}, \tag{19.15}$$

the planner has to acquire costly information. Assuming the planner faces the same exogenous shocks as private agents, if the planner acquires information it is subject to collateral constraints based on the new information. We now define the constrained planner's problem.

DEFINITION 2 (Constrained Planner's Problem): *For each firm with collateral p, a planner chooses the loan size K(p) for production and decides whether or not to acquire information about the firm's collateral to maximize welfare (19.13), subject to collateral constraints (19.14) and information constraints (19.15).*

It is intuitively clear that, without collateral and information constraints the planner would optimally lend $K(p) = K^*$ to each firm, since it is efficient to finance all projects at optimal scale. This is what we referred to above as *unconstrained first best*. It is also intuitively clear, from Figure 19.7, that without information constraints it is optimal for the planner to always avoid information acquisition.

In what follows we first study the economy without aggregate shocks, and show that a planner would like to produce information for a wider range of collateral p than short-lived agents. Then, we study the economy with negative aggregate shocks and show that it may still be optimal for the planner to avoid information production, riding the credit boom even when facing the possibility of collapse.

19.3.1. No Aggregate Shocks

The next proposition shows that, when $\beta > 0$, the planner wants to acquire information for a wider range of beliefs p. Given the planner is constrained by both collateral and information considerations, the only source of inefficiency arises from the myopic behavior of all agents, who consider only the benefits of information for one period and not its potential future costs.

PROPOSITION 7: *The planner's optimal range of information-sensitive beliefs is wider than the decentralized range of information-sensitive beliefs from equation* (19.7). *Specifically, the planner produces information if*

$$(1-\beta\lambda)\frac{\gamma}{qA-1} < pK^* - K(p|II) \tag{19.16}$$

and does not produce information otherwise.

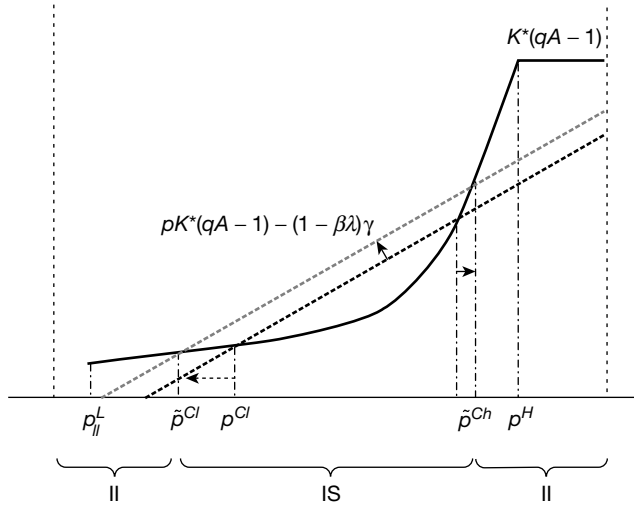


Figure 19.8 Information Acquisition by the Planner

Comparing this condition with equation (19.7), it is clear that the cost of information is effectively lower for the planner. The planner expects to relax collateral constraints if finding out the collateral is good and give a loan to such collateral of K^* in all future periods until a new idiosyncratic shock hits. Decentralized agents, however, do not internalize these future gains when deciding whether to trigger information acquisition or not, since they are myopic and do not weigh the information impact on future generations. This difference widens with the planner discounting $(\beta)$and with the probability that the collateral remains unchanged (λ)

The planner can align incentives easily by subsidizing information production by a fraction $\beta\lambda$ of information acquisition, possibly using lump sum taxes on individuals. In this way, after the subsidy, the cost of information production that agents face is effectively $\gamma\ (1-\beta\lambda)$. Figure 19.8 illustrates this efficiently wider range of information-sensitive beliefs p.

We denote by $\tilde{K}\left(p\right)$ the net effective loan a planner can give a firm with collateral p, considering the effects on future loans and obtained by the upper contour of the solid curve and the upper dashed line of Figure 19.8.

$$\tilde{K}\left(p\right) = \max\left\{K\left(p\mid II\right), pK^*\right\} - \frac{\gamma\ (1-\beta\lambda)}{qA-1},$$

where $K(p\mid II)$ is given in equation (19.5) and the function follows the same schedule as $K(p)$ in equation (19.12) but using instead the effective information cost $\gamma\ (1-\beta\lambda)$ and the cut offs $\tilde{p}^{Ch}$ and $\tilde{p}^{Cl}$ depicted in Figure 19.8.

19.3.2. Aggregate Shocks

In this section we assume that the planner assigns a probability μ per period that a negative shock η will occur at some point in the future. The next proposition shows that there are levels of p for which, even in the presence of the potential future shock, the planner prefers not to produce information, maintaining a high level of current output rather than avoiding a potential reduction in future output. This insight is consistent with the findings of Ranciere, Tonell, and Westermann (2008) who show that *"high growth paths are associated with the undertaking of systemic risk and with the occurrence of occasional crises."*

PROPOSITION 8: *The possibility of a future negative aggregate shock does not necessarily justify acquiring information, reducing current output to avoid potential future crises. In the presence of possible future negative aggregate shocks, the planner produces information if*

$$(1-\beta\lambda)\frac{\gamma}{qA-1} > \frac{(1-\beta\lambda)}{(1-\beta\lambda)+\beta\lambda\mu}\left[pK^* - K\left(p|II\right)\right] + \frac{\beta\lambda\mu}{(1-\beta\lambda)+\beta\lambda\mu}\left[p\tilde{K}(\eta) - \tilde{K}\left(\eta p\right)\right], \quad (19.17)$$

and does not produce information otherwise.

The IS range of beliefs widens if $\left[pK^* - K\left(p\,|II\right)\right] < \left[p\tilde{K}(\eta) - \tilde{K}\left(\eta p\right)\right]$. *Furthermore, the effect of future shocks* η *on the IS range of beliefs increases with their probability* μ.

To build intuition, assume the aggregate shock is not large enough to make $\tilde{K}(\eta) < K^*$ but is large enough to make $\tilde{K}\left(\eta p\right) < K\left(p\,|II\right)$ (for example, $\eta > p^H$ and $p = p^H$). In this case, the aggregate shock, regardless of its probability, does not affect the expected discounted consumption of acquiring information (since even with the shock, a firm with a unit of good land is able to borrow K^*), but the shock reduces the expected discounted consumption of not acquiring information (since with the shock, the loan size declines from $K\left(p\,|II\right)$ to $\tilde{K}\left(\eta p\right)$). In this example, producing information relaxes the potential borrowing constraint in the case of a future negative shock. Hence, when that shock is more likely, there are more incentives to acquire information.

Now assume larger shocks. Take, as an example, the extreme case $\eta = 0$, such that all collateral becomes bad. In this case, condition (19.17) simply becomes

$$(1-\beta\lambda+\beta\lambda\mu)\frac{\gamma}{qA-1} < pK^* - K\left(p\,|II\right),$$

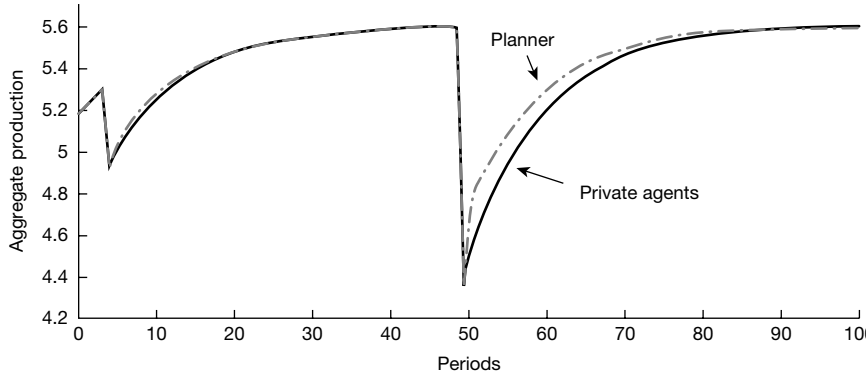


Figure 19.9 Dynamics with an Aggregate Shock $\eta = 0.91$

increasing effective information costs and, hence, reducing the incentives to acquire information. In this extreme case the planner wants to acquire less information than in the absence of shocks (condition 19.16) but still wants to acquire more information than decentralized agents (condition 19.7).

19.3.2.1 Discussion of Dynamics

There are aggregate shocks that induce the same dynamics in the planning and decentralized economies. For example, if $\hat{p} > p_H$ and aggregate shocks are small, then both dynamics are identical to the solid curve in Figure 19.7. In essence the shock does not induce information production in either of the two economies.

There are, however, aggregate shocks that may induce different dynamics between planning and decentralized economies. As an illustration, consider the numerical example in Section IID. If $\beta = 0.9$, then the planner's range for information acquisition is $[0.16, 0.85]$, wider than the decentralized case depicted in Figure 19.3.

Figure 19.9 shows dynamics when aggregate shocks of size $\eta = 0.91$ hit in periods 5 and 50. In this case decentralized agents do not acquire information when the shock hits but the planner does, inducing different dynamics.

The solid curve is identical to the lower dashed curve in Figure 19.5 for the decentralized economy. The dashed curve shows that the planner induces less production in the period of the shock, when acquiring information, but induces a faster recovery afterwards. Since private agents do not value the future, they prefer to produce more in the year of the crisis, not internalizing the costs in terms of a slower recovery. Agents are myopic and do not take into account the effect of their decisions during crises for future generations. This inefficiency is the direct result of our overlapping generations environment and naturally disappears in a dynastic model in which agents value the consumption of future generations.

19.4. CONCLUSIONS

It has been difficult to explain financial crises and how they are linked to credit booms. "Large shocks" or multiple equilibria do not incorporate credit booms and are not convincing explanations of financial crises. Further, they do not lead to policy recommendations. Explaining a financial crisis requires the modeling discipline of fixing the shock size and showing how that shock can sometimes have no effect and sometimes lead to a crisis. Our explanation is based on the endogenous dynamics of information in the economy which creates fragility as a rational credit boom develops. Confidence is lost when a long-lasting credit boom is tipped by a potentially small shock.

The amount of information in an economy is time varying. It is not optimal for lenders to produce information every period about the borrowers because it is costly. In that case, the information about the collateral degrades over time; a kind of amnesia sets in. Instead of knowing which borrowers have good collateral and which have bad collateral, all collateral starts to look alike. These dynamics of information result in a credit boom in which firms with bad collateral start to borrow. During the credit boom, output and consumption rise, but the economy becomes increasingly fragile. The economy becomes more susceptible to small shocks. If information production becomes a credible threat, all collateral with depreciated information can borrow less: a credit crunch. Alternatively, if information is effectively produced after such a shock, firms with bad collateral cannot access credit: a financial crisis.

Why did complex securities, such as subprime mortgage-backed securities, play a leading role in the recent financial crisis? Agents choose (and construct) collateral that has a high perceived quality when information is not produced and collateral that has a high cost of producing information. For example, to maximize borrowing firms will tend to use complex securities linked to land, such as mortgage-backed securities. The opacity and complexity of collateral securities is endogenous, as part of the credit boom. This increases fragility over time.

A credit boom results in output and consumption rising, but it also increases systemic fragility. Consequently, a credit boom presents a delicate problem for regulators and the central bank. We show that a social planner would produce more information than private agents but would not always want to eliminate fragility. Our model matches the main outline of the recent financial crisis. The crisis followed a credit boom in which increasing amounts of complex mortgages were securitized. Short-term debt in the form of repo and asset-backed commercial paper used a variety of securitized debt as collateral, including subprime mortgage-backed securities. This outline of the crisis is more generally a description of historical banking panics, as well, though this is a subject for future research. We focus on exogenous shocks to the expected value of collateral to

trigger crises. However, in Gorton and Ordoñez (2013) we show not only that crises can be triggered by exogenous shocks to productivity but also that they may even arise endogenously as the credit boom grows, without the need for any exogenous shock.

APPENDIX

A. Proof of Proposition 2

Point 1 is a direct consequence of $K\left(p\,|\,\gamma\right)$ being monotonically increasing in p for $p < p^H$ and independent of p for $p > p^H$.

To prove point 2 we derive the function $\hat{K}\left(\gamma\,|p\right)$, which is the inverse of the $K\left(p\,|\,\gamma\right)$, and analyze its properties. Consider first the extreme case in which information acquisition is not possible (or $\gamma = \infty$). In this case the limit to financial constraints is the point at which $K^* = pC$; lenders will not acquire information but will not lend more than the expected value of collateral, pC. Then, the function $\hat{K}\left(\gamma\,|p\right)$ has two parts. One for $p \geq \frac{K^*}{C}$ and the other for $p < \frac{K^*}{C}$.

(i) $p \geq \frac{K^*}{C}$:

$$\hat{K}\left(\gamma\,|p\right) = \begin{cases} K^* & \text{if } \gamma_1^H \leq \gamma \\ \frac{\gamma}{(1-p)(1-q)} & \text{if } \gamma^L \leq \gamma < \gamma_1^H \\ pK^* - \frac{\gamma}{(qA-1)} & \text{if } \gamma < \gamma^L, \end{cases}$$

where γ_1^H comes from equation (19.8). Then

$$\gamma_1^H = K^*\left(1-p\right)\left(1-q\right) \tag{19A.1}$$

and γ^L comes from equation (19.11). Then

$$\gamma^L = pK^*\,\frac{\left(1-p\right)\,\left(1-q\right)\,\left(qA-1\right)}{\left(1-p\right)\,\left(1-q\right)\,+\,\left(qA-1\right)} \tag{19A.2}$$

(ii) $p < \frac{K^*}{C}$:

$$\hat{K}\left(\gamma\,|p\right) = \begin{cases} pC & \text{if } \gamma_2^H \leq \gamma \\ \frac{\gamma}{(1-p)(1-q)} & \text{if } \gamma^L \leq \gamma < \gamma_2^H \\ pK^* - \frac{\gamma}{(qA-1)} & \text{if } \gamma < \gamma^L, \end{cases}$$

where γ_2^H in this region comes from equation (19.9). Then

$$\gamma_2^H = p\left(1 - p\right)\left(1 - q\right)C \qquad \textbf{(19A.3)}$$

and γ^L is the same as above.

It is clear from the function $\hat{K}\left(\gamma \mid p\right)$ that, for a given p, borrowing is independent of γ in the first region, it is increasing in the second region (information-insensitive regime), and it is decreasing in the last region (information-sensitive regime).

B. Proof of Proposition 3

1. $\hat{p}$ is information-sensitive $\left(\hat{p} \in \left[p^{Cl}, p^{Ch}\right]\right)$: In this case, information about the fraction $(1 - \lambda)$ of collateral that gets an idiosyncratic shock is reacquired every period t. Then $f(1) = \lambda\hat{p}$, $f\left(\hat{p}\right) = (1 - \lambda)$ and $f(0) = \lambda\left(1 - \hat{p}\right)$. Considering $K(0) = 0$,

$$W_t^{IS} = \bar{K} + \left[\lambda\hat{p}K(1) + (1 - \lambda)K\left(\hat{p}\right)\right]\left(qA - 1\right). \qquad \textbf{(19B.1)}$$

Aggregate consumption W_t^{IS} does not depend on t; it is constant at the level at which information is reacquired every period.

2. *$\hat{p}$ is information-insensitive* $\left(\hat{p} > p^{Ch} \text{ or } \hat{p} < p^{Cl}\right)$: Information on collateral that suffers an idiosyncratic shock is not reacquired, and at period t, $f(1) = \lambda^t\hat{p}$, $f\left(\hat{p}\right) = \left(1 - \lambda^t\right)$ and $f(0) = \lambda^t\left(1 - \hat{p}\right)$. Since $K(0) = 0$,

$$W_t^{II} = \bar{K} + \left[\lambda^t\hat{p}K(1) + \left(1 - \lambda^t\right)K\left(\hat{p}\right)\right]\left(qA - 1\right). \qquad \textbf{(19B.2)}$$

Since $W_0^{II} = \bar{K} + \hat{p}K(1)\left(qA - 1\right)$ and $\lim_{t\to\infty} W_t^{II} = \bar{K} + K\left(\hat{p}\right)\left(qA - 1\right)$, the evolution of aggregate consumption depends on $\hat{p}$. A credit boom ensues, and aggregate consumption grows over time, whenever $K\left(\hat{p}\right) > \hat{p}K(1)$, or

$$\frac{\gamma}{\left(1 - \hat{p}^*\right)\left(1 - q\right)} > \hat{p}^*K^*.$$

C. Proof of Proposition 4

Assume a negative aggregate shock of size η after t periods without an aggregate shock. Aggregate consumption before the shock is given by equation (19B.2) because we assume $\hat{p} > p^H$ and the average collateral does not induce information. In contrast, aggregate consumption after the shock is

$$W_{t|\eta} = \bar{K} + \left[\lambda^t\hat{p}K(\eta) + \left(1 - \lambda^t\right)K\left(\eta\hat{p}\right)\right]\left(qA - 1\right).$$

Defining the reduction in aggregate consumption as $\Delta(t|\eta) = W_t - W_{t|\eta}$

$$\Delta(t|\eta) = \left[\lambda^t \hat{p}\left[K(1) - K(\eta)\right] + \left(1-\lambda^t\right)\left[K\left(\hat{p}\right) - K\left(\eta\hat{p}\right)\right]\right]\left(qA-1\right).$$

That $\Delta(t|\eta)$ is nondecreasing in η is straightforward. That $\Delta(t|\eta)$ is nondecreasing in t follows from

$$\hat{p}\left[K(1) - K(\eta)\right] \leq \left[K\left(\hat{p}\right) - K\left(\eta\hat{p}\right)\right],$$

which holds because $K\left(\hat{p}\right) = K(1)$ (by assumption $\hat{p} > p^H$), and $K(p)$ is monotonically decreasing in p.

D. Proof of Proposition 5

If the negative shock happens in period t, the belief distribution is $f(\eta) = \lambda^t\hat{p}, f\left(\eta\hat{p}\right) = \left(1-\lambda^t\right)$, and $f(0) = \lambda^t\left(1-\hat{p}\right)$.

In period $t+1$, if information is acquired (IS case), after idiosyncratic shocks are realized, the belief distribution is $f_{IS}(1) = \lambda\eta\hat{p}\left(1-\lambda^t\right)$, $f_{IS}(\eta) = \lambda^{t+1}\hat{p}, f_{IS}\left(\hat{p}\right) = (1-\lambda)$, $f_{IS}(0) = \lambda\left[\left(1-\lambda^t\hat{p}\right) - \eta\hat{p}\left(1-\lambda^t\right)\right]$. Hence, aggregate consumption at $t+1$ in the *IS* scenario is

$$\begin{aligned} W_{t+1}^{IS} = &\bar{K} + \left[\lambda\eta\hat{p}\left(1-\lambda^t\right)K^* + \lambda^{t+1}\hat{p}K(\eta)\right. \\ &\left. + (1-\lambda)K\left(\hat{p}\right)\right]\left(qA-1\right). \end{aligned} \tag{19D.1}$$

In period $t+1$, if information is not acquired (II case), after idiosyncratic shocks are realized, the belief distribution is $f_{II}(\eta) = \lambda^{t+1}\hat{p}, f_{II}\left(\hat{p}\right) = (1-\lambda), f_{II}\left(\eta\hat{p}\right) = \lambda\left(1-\lambda^t\right), f_{II}(0) = \lambda^{t+1}\left(1-\hat{p}\right)$. Hence, aggregate consumption at $t+1$ in the *II* scenario is

$$\begin{aligned} W_{t+1}^{II} = &\bar{K} + \left[\lambda^{t+1}\hat{p}K(\eta) + \lambda\left(1-\lambda^t\right)K\left(\eta\hat{p}\right)\right. \\ &\left. + (1-\lambda)K\left(\hat{p}\right)\right]\left(qA-1\right). \end{aligned} \tag{19D.2}$$

Taking the difference between aggregate consumption at $t+1$ between the two regimes,

$$W_{t+1}^{IS} - W_{t+1}^{II} = \lambda\left(1-\lambda^t\right)\left(qA-1\right)\left[\eta\hat{p}K^* - K\left(\eta\hat{p}\right)\right]. \tag{19D.3}$$

This expression is nonnegative for all $\eta\hat{p}K^* \geq K\left(\eta\hat{p}\right)$, or alternatively, for all $\eta\hat{p} < \overline{\eta\hat{p}} \equiv \frac{1}{2} + \sqrt{\frac{1}{4} - \frac{\gamma}{K^*(1-q)}}$. From equation (19.11), $p^{Ch} < \overline{\eta\hat{p}} < p^H$.

E. Proof of Proposition 6

Assume at period zero that the belief distribution is $f(0) = 1 - \hat{p}$ and $f(1) = \hat{p}$. The original variance of beliefs is

$$\mathrm{Var}_0(p) = \hat{p}^2(1-\hat{p}) + (1-\hat{p})^2\hat{p} = \hat{p}(1-\hat{p}).$$

At period t, during a credit boom, the belief distribution is $f(0) = \lambda^t(1-\hat{p}), f(\hat{p}) = 1 - \lambda^t$, and $f(1) = \lambda^t\hat{p}$. Then, at period t the variance of beliefs is

$$\mathrm{Var}_t(p\,|II) = \lambda^t\left[\hat{p}^2(1-\hat{p}) + (1-\hat{p})^2\hat{p}\right] = \lambda^t\hat{p}(1-\hat{p}),$$

decreasing in the length of the boom t.

Assume a shock η at period t that triggers information acquisition about collateral with belief $\eta\hat{p}$. If the shock is *"small"* $\left(\eta > p^{Ch}\right)$, there is no information acquisition about collateral *known to be good before the shock*. If the shock is *"large"* $\left(\eta < p^{Ch}\right)$, there is information acquisition about collateral *known to be good before the shock*. Now we study these two cases when the shock arises after a credit boom of length t.

1. $\eta > p^{Ch}$. The distribution of beliefs in case information is generated is given by $f(0) = \lambda^t(1-\hat{p}) + \left(1-\lambda^t\right)\left(1-\eta\hat{p}\right), f(\eta) = \lambda^t\hat{p}$, and $f(1) = \left(1-\lambda^t\right)\eta\hat{p}$. Then, at period t the variance of beliefs with information production is

$$\mathrm{Var}_t(p\,|IS) = \lambda^t\hat{p}(1-\hat{p})\eta^2 + \left(1-\lambda^t\right)\eta\hat{p}\left(1-\eta\hat{p}\right).$$

Then

$$\mathrm{Var}_t(p\,|IS) - \mathrm{Var}_t(p\,|II) = \left(1-\lambda^t\right)\eta\hat{p}\left(1-\eta\hat{p}\right) - \lambda^t\hat{p}(1-\hat{p})\left(1-\eta^2\right),$$

increasing in the length of the boom t.

2. $\eta < p^{Ch}$. The distribution of beliefs in case information is produced is given by $f(0) = \lambda^t(1-\hat{p}) + \left(1-\lambda^t(1-\hat{p})\right)\left(1-\eta\hat{p}\right)$, and $f(1) = \left(1-\lambda^t(1-\hat{p})\right)\eta\hat{p}$. Then, at period t the variance of beliefs with information production is

$$\mathrm{Var}_t(p\,|IS) = \lambda^t\hat{p}(1-\hat{p})\eta^2\hat{p} + \left(1-\lambda^t(1-\hat{p})\right)\eta\hat{p}\left(1-\eta\hat{p}\right).$$

Then

$$\mathrm{Var}_t(p\,|IS) - \mathrm{Var}_t(p\,|II) = \left(1-\lambda^t(1-\hat{p})\right)\eta\hat{p}\left(1-\eta\hat{p}\right) - \lambda^t\hat{p}(1-\hat{p})\left(1-\eta^2\hat{p}\right),$$

also increasing in the length of the boom t.

The change in the variance of beliefs also depends on the size of the shock. For very large shocks ($\eta \to 0$) the variance can decline. This decline is lower the larger is t.

F. Proof of Proposition 7

Denote the expected discounted consumption sustained by a unit of collateral with belief p if producing information as $V^{IS}(p)$ and if not producing information as $V^{II}(p)$. The value function from such a unit of land is then $V(p) = \max\left\{V^{IS}(p), V^{II}(p)\right\}$.

If acquiring information, expected discounted consumption is

$$V^{IS}(p) = pK^*(qA-1) - \gamma + \beta\left[\lambda\left(pV(1) + (1-p)V(0)\right) + (1-\lambda)V(\hat{p})\right] + pC.$$

Since we know that for $p = 0$ and $p = 1$ there is no information acquisition, $\left(V(1) = V^{II}(1) \text{ and } V(0) = V^{II}(0)\right)$, and we can compute

$$V(1) = K^*(qA-1) + \beta\left[\lambda V(1) + (1-\lambda)V(\hat{p})\right] + pC,$$

and

$$V(0) = 0 + \beta\left[\lambda V(0) + (1-\lambda)V(\hat{p})\right] + pC.$$

Taking expectations

$$pV(1) + (1-p)V(0) = \frac{pK^*(qA-1)}{1-\beta\lambda} + \frac{\beta(1-\lambda)}{1-\beta\lambda}V(\hat{p}) + \frac{pC}{1-\beta\lambda},$$

and solving for $V^{IS}(p)$, we get

$$V^{IS}(p) = \frac{pK^*(qA-1)}{1-\beta\lambda} - \gamma + Z(p,\hat{p}), \qquad \textbf{(19F.1)}$$

where

$$Z(p,\hat{p}) = \frac{\beta(1-\lambda)}{1-\beta\lambda}V(\hat{p}) + \frac{pC}{1-\beta\lambda}.$$

If *not* acquiring information, expected discounted consumption is

$$V^{II}(p) = K(p|II)(qA-1) + \beta\left[\lambda V(p) + (1-\lambda)V(\hat{p})\right] + pC.$$

Assume $V(p) = V^{II}(p)$, then

$$V^{II}(p) = \frac{K(p|II)(qA-1)}{1-\beta\lambda} + Z(p,\hat{p}), \qquad \textbf{(19F.2)}$$

and $V(p)$ is indeed information-insensitive if $V^{II}\left(p\right) > V^{IS}\left(p\right)$

$$(1-\beta\lambda)\ \frac{\gamma}{qA-1} > pK^{*} - K\left(p\,|II\right).$$

Similarly, assume $V\left(p\right) = V^{IS}\left(p\right)$. We denote as $V^{II}\left(p\,|Dev\right)$ the expected discounted consumption from deviating and not producing information only for one period. Then

$$V^{II}\left(p\,|Dev\right) = K\left(p\,|II\right)\left(qA-1\right) + \beta\left[\lambda V^{IS}\left(p\right) + (1-\lambda)\,V\left(\hat{p}\right)\right] + pC$$

replaces equation (19F.1),

$$V^{II}\left(p\,|Dev\right) = K\left(p\,|II\right)\left(qA-1\right) + \beta\left[\lambda\left(\frac{pK^{*}\left(qA-1\right)}{1-\beta\lambda} - \gamma + Z\left(p,\hat{p}\right)\right)\right.$$
$$\left. + (1-\lambda)\,V\left(\hat{p}\right)\right] + pC,$$

and plugging in $Z\left(p,\hat{p}\right)$ and rearranging, obtain

$$V^{II}\left(p\,|Dev\right) = \left[K\left(p\,|II\right) + \frac{\beta\lambda pK^{*}}{1-\beta\lambda}\right]\left(qA-1\right) - \beta\lambda\gamma + Z\left(p,\hat{p}\right).$$

$V(p)$ is indeed information-sensitive if $V^{II}\left(p\,|Dev\right) < V^{IS}\left(p\right)$, which is again

$$(1-\beta\lambda)\ \frac{\gamma}{qA-1} < pK^{*} - K\left(p\,|II\right).$$

This result effectively means that the decision rule for the planner is the same as the decision rule for decentralized agents, but with $\beta > 0$ for the planner and $\beta = 0$ for the agents.

This result allows us to characterize value functions in equilibrium generally as

$$V\left(p\right) = \frac{\tilde{\pi}\left(p\right)}{1-\beta\lambda} + Z\left(p,\hat{p}\right), \qquad \textbf{(19F.3)}$$

where $\tilde{\pi}\left(p\right) = \tilde{K}\left(p\right)\left(qA-1\right)$ and $\tilde{K}\left(p\right) = \max\{K\left(p\,|II\right), pK^{*} - \frac{\gamma(1-\beta\lambda)}{(qA-1)}\}$, which is the same as array (19.12) but with new cutoffs given by lower effective costs of information $\gamma\ (1-\beta\lambda)$.

G. Proof of Proposition 8

Without loss of generality we assume the negative shock η can happen only once. Until the shock occurs, its ex ante probability is μ per period, turning to zero after the shock is realized. This assumption just simplifies the analysis because,

conditional on a shock, we can impose the results obtained previously without aggregate shocks. Furthermore, we do not need to keep track of all the possible paths of shocks and beliefs. Generalizing this result just requires more algebra but hides the main forces at work behind the results.

Denote by $\hat{V}(p)$ the expected discounted consumption sustained by a unit of collateral with belief p prior to the realization of the shock. As in Proposition 7, denote by $V(p)$ the expected discounted consumption sustained by a unit of collateral with belief p after the shock is realized—hence, in the absence of possible future shocks. This is convenient because we can replace value functions after the shock with the results from Proposition 7 and because we do not need to keep track of different paths of beliefs.

The value of producing information (IS) in periods preceding potential shocks is

$$\begin{aligned}\hat{V}^{IS}(p) = {} & pK^*(qA-1) - \gamma + \beta(1-\mu)\lambda\left[p\hat{V}(1) + (1-p)\hat{V}(0)\right] \\ & + \beta(1-\mu)(1-\lambda)\hat{V}(\hat{p}) + \beta\mu\lambda\left[pV(\eta) + (1-p)V(0)\right] \\ & + \beta\mu(1-\lambda)V(\eta\hat{p}) + pC.\end{aligned}$$

Again we know that for $p=0$ and $p=1$ there is no information acquisition, $\left(\tilde{V}(1) = \tilde{V}^{II}(1) \text{ and } \tilde{V}(0) = \tilde{V}^{II}(0)\right)$ and we can compute

$$\begin{aligned}& p\tilde{V}(1) + (1-p)\tilde{V}(0) = \\ & \frac{1}{1-\beta\lambda(1-\mu)}\left[pK^*(qA-1) + \beta(1-\mu)(1-\lambda)\hat{V}(\hat{p}) + pC\right] \\ & + \frac{1}{1-\beta\lambda(1-\mu)}\left[\beta\mu\lambda\left(pV(\eta) + (1-p)V(0)\right) + \beta\mu(1-\lambda)V(\eta\hat{p})\right].\end{aligned}$$

Also, using value functions in the absence of shocks, $V(p)$, from equation (19F.3):

$$pV(\eta) + (1-p)V(0) = \frac{p\tilde{K}(\eta)(qA-1)}{1-\beta\lambda} + Z(p,\hat{p}).$$

Plugging these results in $\hat{V}^{IS}(p)$ and rearranging we obtain

$$\begin{aligned}\hat{V}^{IS}(p) = {} & \frac{pK^*(qA-1)}{1-\beta\lambda(1-\mu)} - \gamma + \frac{\beta\lambda\mu}{1-\beta\lambda(1-\mu)}\left[\frac{p\tilde{K}(\eta)(qA-1)}{1-\beta\lambda} + Z(p,\hat{p})\right] \\ & + \hat{Z}(p,\hat{p},\eta,\mu), \qquad \textbf{(19G.1)}\end{aligned}$$

where

$$\hat{Z}(p,\,\hat{p},\,\eta,\,\mu) = \frac{\beta(1-\lambda)\left[(1-\mu)\hat{V}(\hat{p}) + \mu\hat{V}(\eta\hat{p})\right] + pC}{1-\beta\lambda(1-\mu)}.$$

The value of NOT producing information (II) in periods preceding potential shocks:

$$\hat{V}^{II}(p) = K(p|II)(qA-1) + \beta(1-\mu)\lambda\hat{V}(\hat{p}) + \beta(1-\mu)(1-\lambda)\hat{V}(\hat{p}) + \beta\mu\lambda V(\eta p) + \beta\mu(1-\lambda)V(\eta\hat{p}) + pC.$$

Assuming $\hat{V}(p) = \hat{V}^{II}(p)$,

$$\hat{V}^{II}(p) = \frac{K(p|II)(qA-1)}{1-\beta\lambda(1-\mu)} + \frac{\beta\lambda\mu}{1-\beta\lambda(1-\mu)}\left[\frac{\tilde{K}(\eta p)(qA-1)}{1-\beta\lambda} + Z(p,\hat{p})\right] + \hat{Z}(p,\hat{p},\eta,\mu), \quad \textbf{(19G.2)}$$

and $\hat{V}(p)$ is indeed information insensitive if $\hat{V}^{II}(p) > \hat{V}^{IS}(p)$, which happens if

$$\frac{\gamma}{(qA-1)}(1-\beta\lambda) < \frac{(1-\beta\lambda)}{(1-\beta\lambda+\beta\lambda\mu)}\left[pK^* - K(p|II)\right] + \frac{\beta\lambda\mu}{(1-\beta\lambda+\beta\lambda\mu)}\left[p\hat{K}(\eta) - \hat{K}(\eta p)\right].$$

Assuming $\hat{V}(p) = \hat{V}^{IS}(p)$, the question is if the planner gains anything by deviating and not producing information for one period. We denote this possibility as $\hat{V}(p|Dev)$

$$\hat{V}^{II}(p|Dev) = K(p|II)(qA-1) + \beta\lambda(1-\mu)\left[\frac{pK^*(qA-1)}{1-\beta\lambda(1-\mu)} - \gamma\right] + \hat{Z}(p,\hat{p},\eta,\mu) + \frac{\beta\lambda\mu}{1-\beta\lambda(1-\mu)}\left[\frac{\tilde{K}(\eta p)(qA-1)}{1-\beta\lambda} + Z(p,\hat{p}) + \beta\lambda(1-\mu)\frac{\tilde{K}(\eta p)(qA-1)}{1-\beta\lambda}\right].$$

$\hat{V}(p)$ is indeed information-insensitive if $\hat{V}^{II}(p|Dev) > \hat{V}^{IS}(p)$, which happens if

$$\frac{\gamma}{(qA-1)}(1-\beta\lambda) < \frac{(1-\beta\lambda)}{(1-\beta\lambda+\beta\lambda\mu)}\left[pK^* - K(p|II)\right] + \frac{\beta\lambda\mu}{(1-\beta\lambda+\beta\lambda\mu)}\left[p\hat{K}(\eta) - \hat{K}(\eta p)\right]$$

which is the same condition obtained before. Based on this condition, the following lemmas are self-evident.

LEMMA 1: *Incentives to acquire information are larger in the presence of future shocks if* $pK^* - K(p|II) < p\tilde{K}(\eta) - \tilde{K}(\eta p)$, *and smaller otherwise. Hence, whether*

there are more or fewer incentives to acquire information in the presence of shocks just depends on their size η, and not on their probability μ.

LEMMA 2: *If in the presence of aggregate shocks there are more incentives to acquire information, these are larger the larger the difference between* $pK^* - K\left(p|II\right)$ *and* $p\tilde{K}(\eta) - \tilde{K}\left(\eta p\right)$ *and the larger* μ.

The first part of the lemma is trivial. The second arises from noting the weight assigned to $p\tilde{K}(\eta) - \tilde{K}\left(\eta p\right)$ increases with μ. These two lemmas, together with the condition for information acquisition we derived, provide a complete characterization of the *IS* and *II* ranges of beliefs under the possibility of a future aggregate shock η that occurs with probability μ, and that is summarized in the proposition.

REFERENCES

Allen, Franklin, and Douglas Gale. 2004. "Financial Fragility, Liquidity, and Asset Prices." *Journal of the European Economic Association* 2 (6): 1015–48.

Andolfatto, David. 2010. "On the Social Cost of Transparency in Monetary Economies." Federal Reserve Bank of St. Louis, Working Paper 2010–001.

Andolfatto, David, Aleksander Berentsen, and Christopher J. Waller. Forthcoming. "Optimal Disclosure Policy and Undue Diligence." *Journal of Economic Theory.*

Bianchi, Javier. 2011. "Overborrowing and Systemic Externalities in the Business Cycle." *American Economic Review* 101 (7): 3400–26.

Bigio, Saki. 2012. "Endogenous Liquidity and the Business Cycle." Unpublished.

Borio, Claudio, and Mathias Drehmann. 2009. "Assessing the Risk of Banking Crises—Revisited." *BIS Quarterly Review:* 29–46.

Campello, Murillo, John R. Graham, and Campbell R. Harvey. 2010. "The Real Effects of Financial Constraints: Evidence from a Financial Crisis." *Journal of Financial Economics* 97 (3): 470–87.

Chari, V.V., Ali Shourideh, and Ariel Zetlin-Jones. 2012. "Collapse of Reputation in Secondary Loan Markets." Unpublished.

Christiano, Lawrence, Roberto Motto, and Massimo Rostagno. Forthcoming. "Risk Shocks." *American Economic Review*.

Claessens, Stijn, M. Ayhan Kose, and Marco E. Terrones. 2011. "Financial Cycles: What? How? When?" International Monetary Fund Working Paper 11/76.

Collyns, Charles V., and Abdelhak S. Senhadji. 2002. "Lending Booms, Real Estate Bubbles and The Asian Crisis." International Monetary Fund Working Paper 02/20.

Dang, Tri Vi, Gary Gorton, and Bengt Holmström. 2013. "Ignorance, Debt and Financial Crises." Unpublished.

Demiroglu, Cem, and Christopher James. 2012. "How Important Is Having Skin in the Game? Originator-Sponsor Affiliation and Losses on Mortgage-Backed Securities." *Review of Financial Studies* 25 (11): 3217–58.

Diamond, Douglas W., and Philip H. Dybvig. 1983. "Bank Runs, Deposit Insurance, and Liquidity." *Journal of Political Economy* 91 (3): 401–19.

Geanakoplos, John. 1996. "Promises Promises." In *The Economy as an Evolving Complex System II,* edited by W.B. Arthur, S. Durlauf, and D. Lane, 285–320. Reading, MA: Addison-Wesley.

Geanakoplos, John. 2010. "The Leverage Cycle." In *National Bureau of Economic Research Macroeconomics Annual 2009.* Vol. 24, edited by Daron Acemoglu, Kenneth Rogoff, and Michael Woodford, 1–65. Chicago: University of Chicago Press.

Geanakoplos, John, and William Zame. 2010. "Collateral Equilibrium." Unpublished.

Gorton, Gary B. 2010. *Slapped by the Invisible Hand: The Panic of 2007.* New York: Oxford University Press.

Gorton, Gary, and Andrew Metrick. 2010. "Haircuts." *Federal Reserve Bank of St Louis Review* 92 (6): 507–19.

Gorton, Gary, and Andrew Metrick. 2012a. "Securitized Banking and the Run on Repo." *Journal of Financial Economics* 104 (3): 425–51.

Gorton, Gary B., and Andrew Metrick. 2012b. "Who Ran on Repo?" Unpublished.

Gorton, Gary, and Guillermo Ordoñez. 2013. "Crises and Productivity in Good Booms and in Bad Booms." Unpublished.

Gorton, Gary, and George Pennacchi. 1990. "Financial Intermediaries and Liquidity Creation." *Journal of Finance* 45 (1): 49–71.

Gorton, Gary, Andrew Metrick, and Lei Xie. 2012. "The Flight from Maturity." Unpublished.

Guerrieri, Veronica, and Robert Shimer. 2012. "Dynamic Adverse Selection: A Theory of Illiquidity, Fire Sales, and Flight to Quality." Unpublished.

Hanson, Samuel G., and Adi Sunderam. 2013. "Are There Too Many Safe Securities? Securitization and the Incentives for Information Production." *Journal of Financial Economics* 108 (3): 565–84.

Ivashina, Victoria, and David Scharfstein. 2010. "Bank Lending during the Financial Crisis of 2008." *Journal of Financial Economics* 97 (3): 319–38.

Jorda, Oscar, Moritz Schularick, and Alan M. Taylor. 2011. "Financial Crises, Credit Booms, and External Imbalances: 140 Years of Lessons." *IMF Economic Review* 59 (2): 340–78.

Kiyotaki, Nobuhiro, and John Moore. 1997. "Credit Cycles." *Journal of Political Economy* 105 (2): 211–48.

Lagunoff, Roger, and Stacey L. Schreft. 1999. "Financial Fragility with Rational And Irrational Exuberance." *Journal of Money, Credit, and Banking* 31 (3): 531–60.

Lorenzoni, Guido. 2008. "Inefficient Credit Booms." *Review of Economic Studies* 75 (3): 809–33.

Mendoza, Enrique G. 2010. "Sudden Stops, Financial Crises, and Leverage." *American Economic Review* 100 (5): 1941–66.

Mendoza, Enrique G., and Marco E. Terrones. 2008. "An Anatomy Of Credit Booms: Evidence From Macro Aggregates And Micro Data." National Bureau of Economic Research Working Paper 14049.

Minsky, Hyman P. 1986. *Stabilizing an Unstable Economy.* New Haven: Yale University Press.

Ordoñez, Guillermo L. Forthcoming. "Fragility of Reputation and Clustering of Risk-Taking." *Theoretical Economics.*

Pagano, Marco, and Paolo Volpin. 2012. "Securitization, Transparency, and Liquidity." *Review of Financial Studies* 25 (8): 2417–53.

Park, Sun Young. 2011. "The Size of the Subprime Shock." Unpublished.

Ranciere, Romain, Aaron Tornell, and Frank Westermann. 2008. "Systemic Crises and Growth." *Quarterly Journal of Economics* 123 (1): 359–406.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly.* Princeton: Princeton University Press.

Schularick, Moritz, and Alan M. Taylor. 2012. "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870–2008." *American Economic Review* 102 (2): 1029–61.

Xie, Lei. 2012. "The Seasons of Money: ABS/MBS Issuance and the Convenience Yield." Unpublished.

20

Some Reflections on the Recent Financial Crisis

GARY B. GORTON*

20.1. INTRODUCTION

Economic development does not result in the elimination of financial crises. The recent financial crisis of 2007–2009 in the United States and Europe shows that market economies, however much they grow and change, are still susceptible to collapse or near-collapse from financial crisis. This is a staggering thought. And it came as a surprise, as financial crises were thought to be things of the past for developed economies, now only occurring in emerging markets. The fact of the 2007–2008 crisis occurrence should give pause to economists. While it may take many years to fully understand the recent crisis, in this essay I offer some preliminary thoughts on crises. I do not review the academic literature, but rather selectively raise some issues and in passing mention some research.

The recent crisis emphasizes that a "crisis" is a distinct, singular, event. It also raises questions about what constitutes bank money, and what is a "bank," and what is the "banking system"? Understanding the crisis has run into problems from a lack of data, leaving researchers in the dark on many important questions. Of course, knowing what data to collect requires an understanding of the crisis. Ironically, if governments and economists knew what data to collect prior to a crisis, they would then likely understand the fragility of the system and could

* Written as a contribution for *Trade, Globalization and Development: Essays in Honor of Kalyan Sanyal*, edited by Sugata Marjit and Rajat Acharya (Springer Verlag; forthcoming). Some of this essay draws from material in my book *Misunderstanding Financial Crises* (Oxford University Press; 2012). I worked at AIG Financial Products as a consultant from 1996–2008. I thank Doug Diamond, Bengt Holmström, Arvind Krishnamurthy, and Guillermo Ordoñez for comments.

possibly avoid a crisis. It seems that a lack of data and the occurrence of a crisis go hand-in-hand. A crisis is a surprise, coming from an unexpected source. As a result, there is little data. There are other inherent difficulties in studying crises. Although crises are perhaps more common than many supposed before the current crisis, still the usable sample size of events is small.

Central to understanding a crisis must be a *concept* of a crisis. A crisis is a systemic event involving an exit from bank debt. It is sudden and unexpected. In the crisis, the banking system is insolvent if not for suspension of convertibility or government and central bank actions. While this is clear, as an empirical matter it has been both easy and hard to define a "crisis." Events are observable, but the whole story is often not observable. Historically, in the face of bank runs, banks suspended convertibility so this could be taken as indicating the outbreak of a crisis, although even this has a number of difficulties. In the modern era, it is much harder to define an event as a crisis and to date the start and the end of a crisis. This is usually because of government and central bank involvement or expectations of such involvement. But, governments usually act late and there are not runs on banks in every case. The upshot is that there is no unanimity on which events are crises, and less unanimity with respect to the start and end dates of crises. This is a manifestation of the lack of data.

What do we know about financial crises? In fact, despite the above difficulties, we know quite a bit about crises. There are a number of stylized facts about crises that have been identified, some rather recently. The stylized facts can help guide the development of models. It should be the case that models incorporate these facts, some of which have to do with the build-up of fragility prior to the crisis and others with the aftereffects, as well as the crisis itself.

First, we know that financial crises occur in all market economies, though sometimes there are long periods of quiet. Crises occur in developed countries, not just emerging markets. The recent financial crisis shows that the financial system can morph in such a way that a crisis can occur after a fairly long period of quiet. The frequency of financial crises historically and internationally strongly suggests that there is a structural or inherent problem with bank debt.

Secondly, we know that crises are exits from bank debt. But, the recent crisis centered on forms of bank debt that are quite different from most (but not all!) previous crises. Can our theories accommodate these other forms of debt? Generating such an event in a model seems harder when the money involved is, for example, sale and repurchase agreements (repo). In this form of money, each "depositor" receives a bond as collateral. There is no common pool of assets on which bank debt holders have a claim. So, strategic considerations about coordinating with other agents do not arise. This is a challenge for theory and raises issues concerning notions of liquidity and collateral, and generally of the design of trading securities—private money.

There are other facts we know about crises. A crisis is a sudden event, a structural break or a regime switch. A crisis is not just a bad outcome in a setting where there is a range of bad outcomes. A crisis is a uniquely bad outcome. Crises are preceded typically by credit booms. They tend to occur at business cycle peaks. They are very costly.

But, there are many things we do not know about crises. We do not know about the dynamics of crises, why agents form suspicions about the value of the assets or collateral backing bank debt. We do not know how agents' expectations of government actions affect the dynamics of crises. We do not know if the preceding credit booms finance productive activity. We do not know what policies can prevent crises, without repressing the banking system. We do not know much about which regulatory systems have been successful at mitigating the occurrence of crises (except by casual observation). We do not know much about how to update measurement systems to detect the buildup of systemic risk.

Overall, the scales tip towards the "do not know" side. In this essay I explore these issues, many of which are areas of ongoing research. The essay proceeds as follows. In Section 20.2 I examine definitions of crisis and outline the difficulties in empirically defining crises. In Section 20.3 I summarize the stylized facts about crises. Section 20.4 is devoted to crisis theory, in particular, the model of Diamond and Dybvig (1983). In Section 20.5 I inquire more generally about the private production of debt securities for transactions and the relation to crises and macroeconomics. Final thoughts are collected in Section 20.6.

20.2. WHAT IS A FINANCIAL CRISIS?

Answering this question is not straightforward either as an empirical matter or as a theoretical matter. In this section I look at practical definitions used for empirical work and in a later section I look at theoretical concepts. One must be informed by the other. We start with a practical definition, which can be used for empirical work.

A "financial crisis" occurs when bank debt holders run on all or many banks demanding that banks convert their (short-term) debt claims into cash to such an extent that this demand for cash cannot be met. The banking system must then be saved by the central bank or the government. Or, if there is no run on banks—or no observed run—then a financial crisis is a situation where there is significant impairment to the banking system, resulting in closures, bailouts, nationalization, blanket insurance guarantees, or other government assistance.

This is a practical definition because its main elements can be observed. Either runs are observed or the interventions are observed. Historically, most bank runs

could be observed (by those outside the banking system). And if a run cannot be observed (as was the case for most observers in the recent crisis), the effects of the run can be observed—bailouts, closures, etc. The magnitude of the event is eventually apparent, and it is deemed a "crisis."

The first part of the definition states the basic points, which we will return to below. It says that the problem is "systemic," that is the banking *system* cannot meet the demands of the debt holders. In this sense, the system is insolvent. This was clear in the recent financial crisis, although the banking system was the shadow banking system not the regulated banking sector. Federal Reserve Chairman Ben Bernanke, in his Financial Crisis Inquiry Commission testimony, noted that of the "13 . . . most important financial institutions in the United States, 12 were at the risk of failure within a period of a week or two" (Bernanke (2010)). The *systemic* aspect should be stressed, as this will play an important role in developing a *concept* of a financial crisis. A "crisis" is not just a bad event.

The first part of the definition refers to "banks" and "bank money" without stating what these terms mean. Until the recent financial crisis, these terms may have seemed clear. "Banks" are regulated firms that issue demand deposits. But, the recent financial crisis illustrates that "banks" and "bank money" changed over the last thirty years. Banks and bank money change their forms through time. Although bank money is typically thought of demand deposits, that was not always the case. Demand deposits developed over time and the extent of their use as money was not understood for many decades (in the 19th century in the U.S.). See Gorton (2012). Bank money takes many other forms, especially historically. Examples of other forms of private money include private bank notes, commercial paper, bankers' acceptances, bills of exchange, and sale and repurchase agreements (repo). Bank debt—private money—is invariably short-term debt issued by certain kinds of firms. The fact that there are other forms of bank money will present some problems for theories, as discussed below.

The second part of the definition relies on observing government actions, taken to address an impending or realized insolvency of the banking system. The government is reacting to a crisis that has already occurred and is causing banks to fail. Often there was a bank run, perhaps a piecemeal run in which the bank debt is attacked over time, an incipient run. Behind this part of the definition a counterfactual is posed. The point is that there would have been a run had it not been for expectations of the government actions. When there is an expectation that the government or central bank will intervene there may be no run on the banks (although in most cases there are runs anyway, though they may come late in the crisis). Events are driven by expectations that the government or central bank will act, but then it may not act in the end, or it acts late. Events then appear chaotic. One need only look at Indonesia during the Asian Crisis to see

an example of this.[1] The financial crisis may also involve a currency crisis as well. The result is that each crisis seems different, special, although at root it is always about bank runs.

Because expectations are unobservable, a practical definition of a crisis—necessary for empirical work—turns on observed bank bailouts or failures. As the recent crisis dramatically illustrated, these events—bailouts and failures—are the result of the crisis, but the crisis—the run—was not observed by those outside the banking system (academics, regulators, the media, the public). Those outside did not observe the run, but only saw the resulting bailouts and failures. These events then are deemed to be the "crisis." This is a mistake. Bailouts and failures are the effects not the causes. Something caused the failures, and this causal factor occurs rather suddenly.

For modern crises, the practical problem is that understanding crises by outsiders relies on observed events such as firm failures or government actions, and government statistics. This problem is manifest in defining and dating crises. In the modern era the determination of whether an event is a crisis, and when it starts and ends, is based on governments' actions because these are readily observable. Boyd, De Nicolò, and Loukoianova (2011) study the four leading classifications and dating of modern crisis events.[2] They show that for many crises the dating of the start and end dates differ quite significantly. There is also some disagreement on which events are crises. Further, they show that the start dates are late.[3] This is because the government actions follow the crisis which has already begun, often in the form of a quiet run (see Gorton (2012)). The dating of the start and the end of a crisis is largely based on contemporary accounts of the crisis, and there is ambiguity.

The economic data available to study crises are usually scanty. Without much data it is hard to do research. Why is collecting data so hard? First, there is the issue of what data to collect, so there must be some prior detailed knowledge of the world in order to know what should be collected. But, outsiders do not know what to collect. They lack the institutional knowledge to know what to collect. So, academics typically focus on the data that are available. Second, even knowing what to collect, there is usually no real way to collect the data. Firms are no help. Similar to the Tennessee Williams play, we must rely on the kindness of

1. See, e.g., Djiwandono (2000) for an eyewitness view of the events in Indonesia. Also, see, for example, Enoch, Baldwin, Frécaut, and Kovanen (2001).

2. These are the classifications of Demirgüç-Kunt and Detragiache (2002, 2005), Caprio and Klingebiel (1996, 1999), Reinhart and Rogoff (2008), and Laeven and Valencia (2008). Laeven and Valencia's database is available at http://www.luclaeven.com/Data.htm.

3. Boyd, De Nicolò, and Loukoianova (2011) use empirical measures of adverse shocks to the banking industry to forecast subsequent government responses. The government responds after the shocks.

(in this case) traders to provide data purely out of a civic duty. This is not a good position for the academe to be in, but there may be no alternative.

The problem of the lack of data cannot be overemphasized. In the recent crisis, many of the most central questions cannot be addressed adequately because of a lack of data. Then the details of the causes of the crisis—the run—and the dynamics of the crisis cannot be formally documented. Instead, research tends to focus on the topics where there are existing data sets, and then the emphasis and attention shifts to those topics. This then distorts our picture of the crisis. Some topics assume enormous importance only because there is data on these topics. Worse still the absence of evidence on other topics is sometimes taken to be evidence of the absence of the importance of these topics, a logical fallacy. As a result, there can be a large gap between anecdotal and eyewitness accounts and what can be more formally documented.

It is easy to see why the empirical study of financial crises is difficult. While crises are frequent in the sense that they occur in all market economies, still the sample size available for econometric study is small and often the relevant data are not available. Historical research can avoid the problem of expectations of government or central bank intervention. I have studied the U.S. National Banking Era, for example, for this reason. But, this presumes that the historical evidence is really about the same type of event as the crises of the modern era. If crises are always about bank runs, then it makes sense to study historical events.

The problem feeds on itself. Without empirical research on crises, theory is unconstrained and will be lacking content. Without theory the notion of a crisis is vague and there is no guide for empirical work. There ends up being no anchor for research. Without addressing these issues, it is hard to make useful policy recommendations. Despite the practical difficulties in empirically identifying crises and their associated timing, we can safely conclude that there are events—"crises"—that are worse economic outcomes than recessions.

20.3. WHAT DO WE KNOW ABOUT FINANCIAL CRISES?

Not enough is known about financial crises. But, I would say that we do know the following facts about financial crises.

1. Financial crises occur in all market economies.
2. Economies can experience long crisis-free periods.
3. Financial crises are sudden and always involve private money (short-term bank debt)—the money markets in the recent crisis.
4. Crises are typically preceded by credit booms.
5. Crises occur at or near business cycle peaks, when the macroeconomy weakens.

6. Recoveries are prolonged following a financial crisis.
7. Financial crises are costly.

The first point is familiar to historians; market economies in different countries have experienced bank runs throughout their histories. But, these experiences vary internationally and over time. One important factor in determining the frequency of crises is the industrial organization of the banking system, in particular whether branching is allowed or prohibited, whether the banking system is a few large banks or many small banks. Also affecting the frequency of crises is the presence or absence of private bank clearinghouses or an effective central bank, and the presence or absence of effective deposit insurance, bank examination and regulation. Based on these factors countries are more or less likely to experience crises. See, e.g., Calomiris and Gorton (1991).

The industrial organization of banking determines the size and structure of the interbank market, which seems to be a critical factor in determining the likelihood of a crisis. For example, in the U.S. in the National Banking Era the regulations and the geographical distribution of economic activity led to "reserve pyramiding," where country banks would deposit their reserves (at interest) with reserve city banks (in large cities), and then they in turn would deposit reserves (at interest) in central reserve city banks (in still larger cities). This intermediation chain, and associated "fictitious reserves," as they were called, induced fragility.[4] This was not the case in England, for example, where the Bank of England's powerful presence was felt. In general, the structure of interbank markets seems very important in affecting the fragility of the system. The structure of the interbank market may also have played a critical role in the recent financial crisis.

But, and this should be stressed, the heterogeneity of countries' crisis experiences should not obscure the central point of the recurring experience of crises. I take this to be one of the main points of Kindleberger (1978, 1993), Reinhart and Rogoff (2009) and Cassis (2011)—crises occur over and over. Laeven and Valencia (2012) count 147 banking crises over the period 1970–2011. And, in particular, developed economies have crises. Reinhart and Rogoff (2008a) note that "for the advanced economies during the full sample, the picture that emerges is one of serial banking crises." Crises in emerging markets have also been frequent, and have some important unique features.[5] Bordo,

4. See, e.g., Mills (1908). The term also refers to the float of checks; see Lockhart (1921a, b), Sprague (1910), and Richardson (2006). There is a theory literature on interbank markets; see, e.g., Rochet and Tirole (1996), Allen and Gale (2000), Freixas, Parigi, and Rochet (2000), and Dasgupta (2004). And, with respect to modern interbank markets, there is also an empirical literature and simulations of interbank exposures; see Upper (2006).

5. For example, see Diaz-Alejandro (1985), Calvo (1995), Kaminsky and Reinhart (1999), and Dornbusch (2001).

Eichengreen, Klingebiel, and Martinez-Peria (2001) look at 120 years, 1880–2000, and argue that the frequency of crises has doubled since 1973. And, Schularick and Taylor (2009, p. 12) note that "the frequency of banking crises in the 1945–71 period was virtually zero; but since 1971 . . . crises became much more frequent." We do not know why this is so.

There is much work to be done to understand the cross-section and time series heterogeneity of crisis experiences internationally and historically. In particular, it is important to understand the cases where no crisis has occurred for a significant period of time, suggesting that some regulatory or central bank framework was effective. One outstanding example of this is the period in England following the Overend, Gurney Crisis of 1866 until 2007. The prolonged stability of the Canadian banking system is another example.[6] And finally, another example is the period in the U.S. from the advent of deposit insurance in 1934 until 2007, a period I have elsewhere called the Quiet Period. Why were there no crises during these periods? This is an important question to answer to be able to design regulations prevent future crises. *Studying the absence of crises is as important as studying crises.*

That crises always involve runs on private bank debt is clear historically, but perhaps less clear in the modern era. Laeven and Valencia (2008) report that 62 percent of the crises in their modern era sample had bank runs. In discussing the counterfactual, related to the definition I gave above, I said that the other crises would have had bank runs had not expectations and subsequent actions of the government and the central bank not stopped the runs. This point is clearly not obvious. But, the accounts of each crisis suggest that this is in fact the case. Here is where eyewitness accounts and contemporary observations of crises are very important. The dynamics of the runs are changed by the existence of a central bank and the government, and in many cases specific policies were adopted that prevented runs, for example, a blanket guarantee on demand deposits. See the discussion in Gorton (2012).

Financial crises are not predictable events although because of credit booms the buildup of fragility is observable. That credit booms often precede financial crises is well-documented, but not well understood. Documentation is provided by Gourinchas, Valdes, and Landerretche (2001), Collyns and Senhadji (2002), Barajas, Dell'Ariccia, and Levchenko (2007), Schularick and Taylor (2009), Reinhart and Rogoff (2009), Borio and Drehmann (2009), Mendoza and Terrones (2008, 2011), Claessens, Kose, and Terrones (2011), and Elekdag and Wu (2011), among others. These studies use different definitions of "credit boom," although the result that crises are best predicted by a "credit boom" seems robust to the definition. Still, this is a bit troubling.

6. See Bordo, Rockoff, and Redish (1994) and Ratnovski and Huang (2009) for discussions of Canada.

Two issues are not really understood. First, although there is some evidence that the credit booms are associated with house price increases, it is not clear more generally what all the credit is being used for. What is the borrowed money being spent on? Secondly, it is not clear that these credit booms are necessarily evils to be avoided. Are the booms supporting productive activity? Fragility builds-up perhaps, but it may also be the case that the credit is supporting productive activity, at least at the start of the boom. We don't know. See Rancière, Tornell, and Westermann (2008) and Gorton and Ordoñez (2012). These are questions for future research.

Financial crises do not happen at random times, but occur near the peak of the business cycle after the credit boom. Gorton (1988) studied the U.S. National Banking Era, 1864–1914, a period during which banking panics regularly occurred, and shows that this is the case. In that study I showed that the arrival of news forecasting a recession resulted in a panic when the news variable exceeded a threshold. The news arrived near business cycle peaks. In the modern era, the results that there are links between financial crises and recessions are similar. For example, Demirgüç-Kunt and Detragiache (1998) examine the period 1980–1994 and "find that low GDP growth, excessively high real interest rates, and high inflation significantly increase the likelihood of systemic problems in our sample" (p. 83). Also see, e.g., Kaminsky and Reinhart (1999).

Historically, economic downturns that involve a financial crisis are worse than the usual downturns. Cerra and Saxena (2008) find that downturns associated with a financial crisis result in output losses of about 7.5 percent of GDP over the subsequent ten years. Reinhart and Rogoff (2009a, b) find that peak-to-trough declines following a crisis average about nine percent. Toujas-Bernaté and Joly (2011) look at 154 countries over 1970–2008 and find long-last output losses; output is reduced by ten percent after eight years. Reinhart and Reinhart (2010) find that GDP growth and housing prices are significantly lower and unemployment higher in the decade following a crisis compared to the decade before. Caballero, Hoshi, and Kashyap (2008) provide empirical evidence on a channel that prolongs crises, in the case of Japan. Also, see Kannan (2010), who looks at industry level data and finds that industries relying more on external finance grow more slowly following the crisis. Related to the aftermath of crises being worse, Jorda, Schularick, and Taylor (2011) show that "more credit-intensive booms tend to be followed by deeper recessions and slower recoveries." But, overall the interaction between financial crises and the business cycle is not clear. The causality is also not clear.[7]

7. There is some interesting work on crises exacerbating downturns. See, for example, Bordo and Haubrich (2009) and Ziebarth (2011).

Crises seem very costly, but these costs are hard to measure. In particular, it is hard to isolate costs that are due to the crisis and not due to the recession that might have occurred even had there not been a crisis. Aside from measures of output loss (relative to trend), there are other measures, such as the net amounts used to resolve bank failures and also fiscal costs. But, the amounts used to bailout banking systems are usually transfers from taxpayers. These transfers may be distortionary and hence costly, but these costs are very hard to measure. Researchers often use the size of the transfer as a proxy. Researchers have tried to address these cost measurement issues in different ways. See, as examples, Laeven and Valencia (2010), Dell'Ariccia, Detragiache, and Rajan (2008), Boyd, Kwak, and Smith (2005), and Hoggarth, Reis, and Saporta (2002). Other costs, such as social, health and psychological costs have not been systematically measured.[8] See Gorton (2012) for a discussion of the costs literature.

These stylized facts provide some broad guidance for a theory of crises. To be clear, financial crises are bank runs, though the form of the "banks" and the "bank money" changes. Bank debt is vulnerable to runs, and crises are usually an integral part of the business cycle in market economies. The facts are not consistent with crises being caused by distortions from government policies, which may be important but which cannot be the basis for a theory of crises. Government actions to prevent crises or to save the banking system in a crisis may be problematic, but they are responses to possible crises, effects, not causes. The stylized facts require explaining the credit boom prior to the crisis as well as the subsequent prolonged below average recoveries. And, it is important to explain why economies can have long periods of quiet, perhaps due to the success of laws and regulations.

But, as I mentioned in the Introduction, there is much we do not know about crises. We do not know the details of how crises are triggered, or what happens during a crisis to exacerbate or allay agents' fears. We do not really know what policies prevent crises. We do not know much about credit booms, how they get started, why they persist, how they end. We do not know how, or if, credit booms are related to asset price increases. We do not know the links between crises and business cycles.

20.3.1. Crisis Theory

Theoretically, a financial crisis is defined by two essential points. First, a crisis is a singular event. It is a rending, a sundering, or a rupturing, of the normal state of affairs in money markets. A financial crisis is *not* the worst outcome on a continuum of bad events. There is no continuum in an important sense. There

8. Though see Furceri and Zdzienicka (2009) on the effects of crises on human capital.

are booms and recessions, and then there are crises. A crisis is a distinct event. Something happens to make a crisis fundamentally different from the usual economic downturn. Second, while each crisis has important unique features, crises have a common root cause. There is a structural feature of bank debt that makes the debt vulnerable to runs. And the bank debt in question is not just demand deposits. Financial crises are always about bank runs. The bank runs either occur or would have occurred had the government or central bank not intervened or been expected to intervene.

The first point says that a "crisis" is not simply a particularly "bad state" of the world. A crisis is fundamentally different, a different regime. There are normal non-crisis states and there is an extraordinary crisis state. This is why Anna Schwartz (2007) said that "a decline in asset prices of equity stocks, real estate, commodities; depreciation of the exchange value of a national currency; financial distress of a large non-financial firm, a large municipality, a financial industry, or sovereign debtors—are pseudo-financial crises" (p. 245). They may be bad events, wealth may be destroyed or cleanup costs high, but they are not crises. A financial crisis is a *systemic* event. The entire financial system is engulfed. The failure of a large firm or problems in one sector, e.g., savings and loans or the auto industry, are not crises in this sense.

Financial crises repeatedly occur in market economies. The second point is that there is a reason for this. There is a root cause. Agents in the economy need private money to transact. But, this money is vulnerable to runs. Bank runs are crises. Financial crises are caused by bank runs.

The root of the financial crisis problem was elegantly identified by Diamond and Dybvig (1983).[9] Diamond and Dybvig studied a setting where banks must use long-term collateral to back demand deposits. Agents need the demand deposits because of potential shorter-term liquidity needs. The investments are "long" in the sense that if they are liquidated early there is a very low return. "Long" also means relative to the required frequency of agents' transactions for consumption or other short-term needs. The agents need demand deposits to smooth consumption, which is uncertain as some agents may want to consume early. An essential feature of the model is that the interest rate offered on the demand deposits to achieve this smoothing is such that if all agents want to consume early (by withdrawing from the bank), then the bank cannot satisfy these demands. This is the critical fragility in the economy.

A very important point is that there is no way around this basic horizon problem in any market economy. People eat lunch every day, but it takes a long time to build a factory and produce output. People need to pay for their lunch before

9. There is a large literature on the Diamond and Dybvig model, many extensions and discussions, but I will, for the most part, not go into this literature.

the output is realized. This timing is fundamental. Bank debt used for transactions can only be backed by these long investments (which have a low return if liquidated early). The private sector cannot produce riskless assets. These basic facts mean that financial intermediaries will always be involved in "maturity transformation," a term which just restates this fact.

"Maturity transformation" is not a choice. It can't be regulated away. It is inherent in any economy which produces private bank money, that is, any market economy. It is a fundamental fact. Bank money can only be backed by longer-term investments. As we will see later, agents in the economy will strive mightily to design bank debt to overcome this problem. But, without the government, bank debt will always be vulnerable.

Uncertainty about consumption timing is a risk the agents want to shed using bank debt. The problem of long-term collateral backing bank debt is a necessary but not a sufficient condition for crises to occur. To get a crisis—a bank run, Diamond and Dybvig introduce a source of uncertainty that is quite special. It is the uncertainty that each individual bank depositor faces about the actions of other deposit holders. Depositors care about the actions of other depositors if there is a common pool of assets on which they all have pari passu claims—the bank's assets—but the claims are honored sequentially (so they are not in fact pari passu). Note that the assumption of sequential service means that the payout of the bank to an individual depositor depends on the actions of the other depositors. How much a depositor gets back depends on his place in the line. In this setting, depositors may run if they think other depositors are going to run. Each depositor has an incentive to be first in line to withdraw at the bank if he believes that other depositors are going to line up. Beliefs about other depositors' beliefs must depend on something and in Diamond and Dybvig beliefs are coordinated by an extraneous random signal, a sunspot.

The bank run, due to the beliefs coordination problem, displays the second essential condition of the definition of a crisis, discussed above. A run in the Diamond and Dybvig model is fundamentally different from the normal state of affairs. There are no "small" crises in the Diamond and Dybvig model. There are two outcomes: no crisis and crisis. The crisis is a distinct, very different, event. This is consistent with the empirical evidence that there are distinct events that can be called "crises" and which are clearly much worse outcomes than recessions or Anna Schwartz's pseudo-financial crises. The model lays out a convincing setting and shows that the outcome can be very different than the normal state of affairs, a run can occur—a crisis. This was the first model that displayed the two essential points articulated above. In this sense, it provides a coherent picture of a financial crisis.

But, as a theory of crises, the Diamond and Dybvig model is not completely satisfactory. The very phenomenon we want to explain, *why* there is a loss of confidence, is not explained—it is "sunspots." That is, each agent believes that

the other agents will run when they observe "sunspots." While the coordination device is called "sunspots," this is just a name for the multiple equilibria that can occur in the model. There is no explanation for why the economy switches from one equilibrium to another.

The issue of belief coordination is especially troublesome. There is no explanation for why a run would suddenly occur. And so, no empirical predictions or policy implications follow. The empirical evidence shows that financial crises are preceded by credit booms and are related to the business cycle, and that agents are prone to run when public information arrives forecasting a recession. The link between the preceding credit boom and the business cycle provides the structure for belief formation.

Economists have attempted to address the issue of belief formation in the Diamond and Dybvig model (and other similar models). Using the global games approach of Carlsson and van Damme (1993), if some noise and asymmetric information are added to the model the multiplicity can be eliminated or reduced. If each depositor privately observes a signal about the future value of the banks' assets, then the equilibrium can be unique if their private signals about the banks' assets are sufficiently accurate. In this way, the belief coordination problem can be linked to economic fundamentals. There is still a threshold effect, so a crisis is a distinct event.[10] Coordination games can generate large changes in agents' behavior without large changes in economic fundamentals. Agents change their beliefs about the actions of other agents and this can have a large effect. This is a general statement which applies to many phenomena, as long as they can be modeled as a coordination game, where the payoff to any one agent depends on the actions of other agents.

It is important to note that this is a purely formal fix-up to a vexing problem arising in the Diamond and Dybvig model. It can't be tested; no one has every articulated the nature of the private information that bank debtholders might realistically have learned. Kelley and Ó Gráda (2000) and Ó Gráda and White (2003) study the details of who ran on the Emigrant Industrial Savings Bank in 1854 and 1857. It is hard to see what the nature or role of the alleged private information.

There are still more fundamental problems. First, the issue of belief coordination only arises for some forms of bank money. Demand deposits are claims on a common pool of assets—the bank's portfolio of loans, the case where belief coordination arises as a problem. Other forms of bank money can differ from the Diamond and Dybvig model in important ways. There may be a maturity date on the claim, even if it is a short maturity, and there may be no common

10. The important papers are Morris and Shin (2001) and Goldstein and Pauzner (2005). The multiplicity of equilibria can be eliminated in other ways; see, e.g., Postlewaite and Vives (1987).

pool problem. If a "depositor" does not have a claim on a common pool of bank assets, then the actions of other depositors are irrelevant; beliefs about other agents' beliefs then do not matter. Or, if there is no sequential service, no lining up, then claims really are pari passu. But, financial crises are not just about demand deposits. All forms of bank money are vulnerable.

Bank money is short-term debt. The critical feature of bank money is that it retains value so that it can act as a short-term store of value or such that other agents unquestioningly accept it in a transaction, without suspicion of private information held by the counterparty. Bills of exchange and negotiable instruments generally are bank money. This includes private bank notes, commercial paper, bankers' acceptances, money market funds, sale and repurchase agreements, and sight drafts. In fact, the history and evolution of various forms of bank money is rich and complicated. There are many kinds of bank money. See, e.g., Usher (1914), DeRosa (2001), and Ferderer (2003). Longer term bank debt that by design resembles government debt may also be included, that is, securitizations.

Checking accounts have not always been the primary form of bank money, and even today checks are being replaced by ATM machines and on-line banking. See Quinn and Roberds (2008). The issue of whether all forms of bank money are vulnerable to runs was brought to the fore by the recent crisis. The recent financial crisis was not a case of household depositors running on banks. It involved firms, financial and nonfinancial, foreign and domestic, running on shadow banks in the repo and asset-backed commercial paper markets. And, even this type of wholesale run is not new. See Quinn and Roberds (2012) and Schnabel and Shin (2004) who study a run in the wholesale market in Amsterdam in 1763. And, see Flandreau and Ugolini (2011) on the Overend-Gurney Panic of 1866 in England. It seems clear that runs have occurred under a variety of bank money forms.

One of the most important forms of bank money historically was private bank notes. Private bank notes were issued by banks in many countries. Schuler (1992) finds sixty cases of such free banking in history. In some cases these notes were claims on a common pool of assets and in some cases they were not. In the U.S. under state free banking laws banks were required to back their notes with state bonds. In the case of a bank failure—an inability to honor requests for cash from noteholders—the state bonds would be sold (by the state government) and the note holders paid off pro rata. Note holders were paid off pro rata, so there was no common pool problem. Yet, there was a run on banks (banknotes and deposits) during the Panic of 1857.

The recent financial crisis centered on sale and repurchase agreements (repo).[11] In a sale and repurchase agreement (a repo) one party lends/deposits

11. See Gorton (2010) and Gorton and Metrick (2012).

money typically overnight at interest and this depositor receives a specific bond as collateral from the bank borrower. The lender/depositor must return the collateral at the maturity of the repo contract. There is no common pool of assets upon which the "depositor" has a claim.[12] If the borrower/bank fails, then the lender/depositor can unilaterally terminate the contract and sell the collateral. Of course, a depositor need not renew the loan, and will not if there are concerns about the joint event of (1) the solvency of the bank and (2) the value of the collateral.

Repo and free banknotes are two examples of bank money where there is no common pool problem. Demand deposits and asset-backed commercial paper are examples where there is a common pool problem; these forms of bank debt are backed by a common portfolio of assets. We observe runs on both forms of bank money, suggesting that the common pool problem is not the inherent vulnerability.

Another special feature of the Diamond and Dybvig model is the fact that agents do not actually meet and trade, so there are no prices in the model.[13] In the model, terms are set on the bank contracts initially and there are no subsequent prices because there is no subsequent trading among agents. In reality there are two complications. First, with many forms of bank money, including demand deposits and private banknotes, agents directly transact. One agent meets and, for example, writes a check to another agent in exchange for goods. Second, other forms of bank money have maturities; agents do not have the contractual right to withdraw any time.

In the Diamond and Dybvig model, once the agents have deposited money in the bank, there are no later transactions between depositing agents in the model. Some agents, perhaps all agents, go to the bank to withdraw prior to the realization of the investment payoffs. But, they do not transact directly with each other at some price, the price of goods in terms of the bank money. So, there are no prices in the model at the date when agents form beliefs about the actions of other agents.

But, in reality, agents do meet and trade goods or services for bank money. Before the U.S. Civil War when agents transacted they used private bank notes, the liabilities of banks denominated as money (i.e., one dollar bills, five dollar bills, etc.). An agent would go to the store and offer to buy goods with these notes. But, these notes did not trade at par. There was an exchange rate between the notes and gold. That is, there was a price. And prices contain information. It could be that one agent writes a check to another agent, for example. In this case, the relative price of the bank money in terms of goods plays a role, as in other

12. Although see Martin, Skie, and von Thadden (2010).

13. Jacklin (1987) discusses some of the trading restrictions in the Diamond and Dybvig model.

markets. With demand deposits the price is usually par, except in a crisis when checks were discounted.

There are two cases. First, suppose there is a common pool problem. What is the effect of prices? Atkeson (2001) raises this point. In this case of the coordination problem, it is not clear that the multiplicity of equilibria disappears when prices are introduced. Economists have tried to address this issue and in related settings have found that the multiple equilibria remain in the presence of prices. See, e.g., Angeletos and Werning (2006) and Hellwig, Mukherji, and Tsyvinski (2006). We would like to have a detailed theory of how beliefs are formed. This is an ongoing area of research.

The second case occurs when there is no common pool problem. There is no common pool problem in repo, for example. In a repo transaction there is a depositor who lends money and a bank borrower. The depositor receives interest on the loan, which is usually overnight. And the borrower delivers collateral to the depositor, which must be returned when the transaction matures. The collateral is sometimes "haircut," which means that the depositor lends less money than the market value of the collateral provided. For example, $90 million is lent and the collateral is worth $100 million at market prices. In repo, haircuts and interest rates depend on the identity of the counterparty if the collateral is private bonds. Even in an over-the-counter market, at any moment, agents in the market (eventually) know these prices. These prices are formed somehow and are related to agents' beliefs.

Another issue concerns how a crisis ends. If agents run on banks because they believe other agents will run, or because fundamentals have deteriorated, how does the crisis end? It is clearest to think of this before there is a central bank, say during the National Banking Era in the U.S. The run starts—for some reason, time passes, and then agents no longer want to run. Somehow agents' anxiety is assuaged, their beliefs are revised. But, we don't know how this happens.[14] If the government or the central bank takes actions, then agents may revise their beliefs about whatever it was that caused them to run to start with. The details of what this means and how it happens are unclear. Before the Federal Reserve System was in existence, this puzzle is clearer. A run would start, usually in New York City, and banks would suspend convertibility. What happened during the period of suspension that allowed bank to resume convertibility? A model which can explain how a "loss of confidence" occurs needs also to explain how confidence is recovered. Clearly, a model with multiple equilibria as the "explanation" for a crisis has difficulties here.[15]

14. We know that the clearinghouses acted during crises, but we do not know how agents' beliefs were revised in response. We just know that eventually suspension of convertibility was lifted.

15. That is, a "reverse" sunspot just compounds the problem of a lack of an explanation.

The Diamond and Dybvig setting is compelling. Private agents cannot produce debt that is invulnerable to runs. Only long-term private assets are available to back bank debt, which is needed to facilitate shorter-term transactions that some agents need to make to smooth consumption. But, the bank debt is vulnerable. And a crisis in Diamond and Dybvig is a distinct event. Building on Diamond and Dybvig requires a model in which a state of the world occurs causing everyone to run.[16] Clearly, there is much work to be done. Incorporating credit booms into a crisis theory, explaining why there is an association between crises and prolonged recoveries, and explaining how a crisis ends, are all open questions.

20.3.2. Bank Debt

Let's take a step back and ask a general question: why is bank debt used for transactions? Agents could issue their own money. Or firms could issue money. In principle, the "money" could be equity or debt, or indeed, any security. Many such securities are traded in markets that are often described as "liquid." So, a basic question is why bank debt is used as money. Why banks? And why debt?

These questions are related to the notion of "liquidity," a term that is used in different ways in the economics literature. A central contribution of Diamond and Dybvig is their notion of "liquidity" as consumption smoothing. But, there is another notion of liquidity, a quite natural one first articulated by Keynes and similar to traders' intuitive notions. Keynes wrote that an asset is liquid if its value is "more certainly realizable at short notice without loss" (Keynes (1930, p. 67)).

Looking back to the Free Banking Era in the U.S. before the Civil War, one can get a sense of this notion of liquidity. Bank notes traded at discounts from par when the transaction was taking place at any distance from the issuing bank. The discount was uncertain and was determined in informal banknote markets where note brokers made markets and traded. The prices in these markets were reported in newspapers called "banknote reporters" that listed the discounts from par at particular locations. In Philadelphia for example, the banknote reporter would list the discounts on hundreds of notes. For example, a merchant arriving in Philadelphia from Savannah might be carrying the banknotes of a New Orleans bank. New Orleans is a quite a distance from Philadelphia and, depending on the year in which the transaction is taking place, it might

16. There are other models of runs, as well. Diamond and Rajan (2001) show a model of bank fragility that is different than Diamond and Dybvig. It connects the asset side of banks to the liability side more specifically, showing that a kind of fragility is required, and displays a collective action problem. Another interesting example is Rochet and Vives (2004).

have taken a week to ten days to get from Philadelphia to New Orleans. The discount on the note reflected this distance. Discounts were higher for more distant banks. In studying this market I showed that the discounts were not chaotic but rational.[17] But still transacting with banknotes was a problem because the discount had to be determined in a market and recorded by the banknote reporter. Then the banknote reporter had to be consulted, arguments ensued, and the less informed party with weak bargaining power was possibly cheated. The pre-Civil War era is replete with constant complaints about bank notes.

Checks became more prevalent starting in the 1850s and by the 1890s were the dominant form of bank money in the U.S. The transition from bank notes to checks is a very important example of the change in the form of bank money. Demand deposits led to the system of "clearing," the process by which bank checks were returned to the bank where the depositor had an account. In the clearing process this bank would then honor the claim. With many banks, clearing in one location—the clearinghouse—netting of the claims could be accomplished.

It is important to understand that checks didn't exist then (or now) as a widespread form of money without private bank clearinghouses. Clearinghouses are inherent in demand deposits; they were part of the process which allowed checks to be efficient. Since checks must be "cleared" banks face enormous counterparty risk. In the clearing process, a bank may have a large positive net position with another bank. If that bank fails, then it could be disaster. Checks imply clearing, and clearing implies large counterparty exposures on a daily basis. This is the basis for clearinghouse to assume a monitoring and information production role. It makes no sense to think of checks without also thinking of clearinghouses.

Clearing *internalized* the note market. It allowed banks to monitor each other and created incentives to do so. The process of clearing in private bank clearinghouses meant that bank could enforce a price of par on in-state checks.[18] This was accomplished by clearinghouse rules and regulations.

The information environment was fundamentally altered by the role of the clearinghouse.[19] As a result, checks were more liquid than bank notes. With checks the problems of transacting were eased. Of course, the person's identity had to be checked, so transactions still took some time. But, the clearinghouse created liquidity and checks came to dominate private bank notes. The "liquidity" of checks was greater than that of bank notes.

17. See Gorton (1996, 1999).

18. Young (1910, p. 608) writes that the organization could expel weak banks, enabling "the clearing house as a body to exercise such supervision of any weak bank as to amount to a virtual taking over of its management till it is again in sound condition."

19. See Cannon (1910), Gorton (1984, 1985), Timberlake (1984), Gorton and Mullineaux (1987), Richardson (2006), and Moen and Tallman (2010).

To stress the point, there can be no model of demand deposits without including clearinghouses. The clearinghouses ensured that checks traded at par. There were no discounts, as with banknotes. Note that this is important when agents meet and trade, suggesting that such trades should be included in a model.

Gorton and Pennacchi (1993) argue that banks exist to create trading securities that allow for transactions to be "more certainly realizable at short notice without loss," that is to trade a par without suspicions of counterparties or the backing assets of the checks. In particular, a holder of the security need not fear a loss of value to better informed parties when there is a transaction because the security is riskless. There can be no losses to better informed parties. But, Gorton and Pennacchi, like Diamond and Dybvig, did not explain why debt is the security banks issue for transactions. See Holmström (2008).

In Diamond and Dybvig the bank exists to smooth consumption, and in Gorton and Pennacchi the bank exists to produce a trading security that can be used without fear of loss to better informed traders. But, there remains the question of why these securities are debt. Existing theories of debt are not concerned with trading. They explain the existence of debt in settings focused on controlling the corporation, getting repaid when investing in a firm. The setting there is one in which the corporation has private information and the firm's output is not observable or not verifiable. There is no trade beyond the initial investment.

Holmström (2011) and Dang, Gorton, and Holmström (2012) provide a theory of debt as trading securities. They argue that debt is the optimal security for trading because it minimizes the incentive for a counterparty to produce private information about the payoff on the trading security. Adverse selection when transacting can then be avoided (most of the time). Riskless securities cannot be produced by the private sector. But, if agents can only produce information at a cost, then liquid securities are those which reduce the benefits of producing such information. Roughly speaking, debt minimizes the incentive to produce information because it has a bounded upside and that bound can be set as tight as possible by providing the debt holder with the maximum amount in the case of bankruptcy (the 45 degree line in case default occurs).

The debt is "information-insensitive" in two senses. It is immune to the counterparty producing private information in most states of the world, thus avoiding adverse selection. And, secondly, it retains the most value in the face of public information. But, such debt can sometimes become "information-sensitive." Dang, Gorton, and Holmström show that in the case of public bad news, it can be the case that a counterparty in a transaction finds it optimal to produce private information in which case the debt holder must accept adverse selection or trade at a price that is below the conditional expected value of the debt. These are instances of a crisis. A crisis displays the regime switch feature that I discussed above. There is a switch from information-insensitive debt to information-sensitive debt which then causes a collapse of trade.

Gorton and Ordoñez (2011) embed this idea of information-insensitive debt in a dynamic macroeconomic setting and show that a credit boom can occur when agents find that information-insensitive debt is optimal. Over time more and more borrowing occurs because agents "forget" which collateral is high quality. Agents act as if most collateral is the average value, relatively high quality, and make loans on this basis. As the boom proceeds, a "small" shock can cause a switch to information-sensitive debt. A shock which would have no effect early on has a large effect when the boom has been ongoing for some time. The crisis is a sudden regime switch.

This is in contrast to models which display amplification or persistence—important effects to be sure, but which cannot display a crisis in the sense of a sudden regime switch. For example, in the model of Kiyotaki and Moore (1997) a shock is magnified via a feedback effect on the value of collateral. But, every shock, big or small, causes some feedback. There is a continuum of outcomes for a range of shocks, and so, in this setting, a crisis must be a large shock. Similarly, in Bernanke and Gertler (1989) a shock, any shock, creates persistence through reducing the net worth of firms resulting in lower borrowing and lower output. But, every shock results in this effect.[20] My point is that these models cannot produce crises except via a "large" shock. Since the large shock is exogenous, this is not a theory of crises.[21] On the contrary, Dang, Gorton, and Holmström show that fragility is endogenous, via the creation of debt that is information-insensitive. Gorton and Ordoñez (2011) show how a credit boom can endogenously create fragility; a large shock is not required for a crisis.

A woman cannot be a little bit pregnant or a person a little bit dead. There is a crisis or there is not a crisis. This is an important point from Diamond and Dybvig. In Dang, Gorton, and Holmström the crisis occurs when privately-produced money endogenously becomes subject to adverse selection and loses its liquidity. Collateral that is information-insensitive is very hard information. And is the basis of private bank money. The crisis occurs when the collateral is no longer above suspicion, so to speak. The switch from information-insensitive to information-sensitive is the loss of "confidence" and corresponds to the regime switch. Holmström (2011) draws a number of other important implications from these ideas.

This model of debt and associated crisis is very different from the "frictions" incorporated into macro models.[22] Simply put, these models do not generate

20. Also see Bernanke, Gertler, and Gilchrist (1999).

21. And, to be clear, the authors of these papers never claimed that their models were such crisis theories. Others have made this claim since the financial crisis.

22. See Bunnermeier, Eisenbach, and Sannikov (2012) for a survey of macro frictions.

crises. Kiyotaki and Moore (1997) and Bernanke and Gertler (1989) are now—since the crisis—cited, ex post, as examples of the attention paid to financial frictions in macroeconomics. But, these models were not part of the formal modeling approach used in policy circles. Models addressing issues of the persistence of temporary shocks and the amplification of shocks are important. But, they cannot display crises. A macro model that can display a financial crisis is a distinct undertaking from a model which displays persistence of temporary shocks, real effects shocks to net worth, or from other financial frictions. As emphasized above, a crisis is a singular event, not the result of a large shock.

The notion of "frictions" arises when the benchmark model, the neoclassical growth model and complete markets cannot replicate important features of reality. In order to induce this model to replicate various features of reality one then adds "frictions." There is a great deal of discretion here in modeling. The researcher chooses from a smorgasbord of "frictions" to add in order to obtain the desired "result." The problem really is that the benchmark model misses the fact that private money is inherent in market economies. This was first noted a long time ago, for example, by Martin Shubik (1975), but the current crisis strongly suggests that this approach has reached a dead end.

That bank debt is vulnerable to runs in market economies is a fact, like demand curves sloping downward. It is not a "friction" in that sense, but a fundamental feature of market economies. Once again, it is clear that there is much research to do. There are a number of (to me, anyway) exciting directions that are developing in macroeconomics. Examples include Brunnermeier and Sannikov (2010), He and Krishnamurthy (2012), and Maggiori (2012). These models incorporate financial sectors and do not focus on steady states. That is, they do not focus on linearized system dynamics around the steady state. So, they can display crisis-like behavior. On the other hand, while they incorporate financial sectors, the crisis is a big shock. The dynamics are triggered by a large shock which reduces the capital of banks, causing them to have to sell assets. While this may be viewed as a reduced form for a bank run, it is not, in fact, a run. Also see Boissay, Collard, and Smets (2012).

20.3.3. Final Thoughts

President Obama's chief of staff Rahm Emanuel observed during the crisis that: "You never want a serious crisis to go to waste," meaning that it is an opportunity to address long overdue problems in a major way. This is good advice for economists as well. The crisis revives old issues and raises new issues. The human toll from the crisis means that this is quite an urgent task. In order to address these issues documenting what happened during the recent financial crisis is critical to our understanding and remains the first task.

The recent crisis emphasizes a number of points. These are worth repeating. First, the recent crisis was a bank run, in the money markets. Secondly, the recent crisis emphasizes that a financial crisis is a distinct, regime switch-type, event. It was clearly different, worse, larger, than usual recessions. Thirdly, it showed (again) that crises recur in market economies. Fourthly, the crisis also showed that bank money without the common pool problem is vulnerable to runs. Fifth, it poses the question of why crises do not occur during certain periods. What regulation was successful? Sixth, the fact that basic institutions in the economy—banks, bank money—could transform largely without notice, means that our measurement systems are suspect. These are important lessons.

The first two points are the core of the *concept* of a crisis, while the third point emphasizes the fundamental nature of crises in market economies. The theory of crises needs to address the fourth point because, as an empirical matter, all forms of bank money are vulnerable. We know little about why there are long periods of quiet, about what bank regulations are effective or whether it was just good luck that produced these periods. Finally, producing measurement systems that keep up with change are paramount.

I have emphasized that empirical documentation of the crisis is critical, and that it is difficult for outsiders who did not *see* the crisis to know what to document. Finding data is hard, but crucial. Theory cannot be built on newspaper stories.

REFERENCES

Allen, Franklin and Douglas Gale (2000), "Financial Contagion," *Journal of Political Economy* 108, 1–33.

Angeletos, George-Marios and Ivan Werning (2006), "Crises and Prices: Information Aggregation, Multiplicity, and Volatility," *American Economic Review* 96 (5), 1720–36.

Atkeson, Andrew G. (2001), "Rethinking Multiple Equilibria in Macroeconomic Modeling: Comment," in NBER Macroeconomics Annual 2000, ed. Ben Bernanke and Kenneth Rogoff (MIT Press; Cambridge, MA), 162–171.

Barajas, Adolfo, Giovanni Dell'Ariccia, and Andrei Levchenko (2007), "Credit Booms: The Good, the Bad, and the Ugly," working paper.

Bernanke, Ben (2010), "Causes of the Recent Financial and Economic Crisis, " Statement by Ben S. Bernanke, Chairman, Board of Governors of the Federal Reserve System, before the Financial Crisis Inquiry Commission, Washington D.C. (September 2, 2010); see http://www.federalreserve.gov/newsevents/testimony/bernanke20100902a.htm.

Bernanke, Ben and Mark Gertler (1989), "Agency Costs, Net Worth, and Business Fluctuations," *American Economic Review* 79, 14–31.

Bernanke, Ben, Mark Gertler and Simon Gilchrist (1989), "The Financial Accelerator in a Quantitative Business Cycle Framework," in John Taylor and Michael Woodford, eds., Handbook of Macroeconomics (Elsevier Science, North Holland; Amsterdam).

Boissay, Frédéris, Fabrice Colard, and Frank Smets (2012), "Booms and Systemic Banking Crises," European Central Bank, working paper.

Bordo, Michael and Joseph Haubrich (2009), "Credit Crises, Money, and Contractions: A Historical View," Federal Reserve Bank of Cleveland Working Paper No. 09–08.

Bordo, Michael, Barry Eichengreen, Daniela Klingbiel, and Maria Soledad Martinez-Peria (2001), "Is the Crisis Problem Growing More Severe?," *Economic Policy* 16, 51–82.

Bordo, Michael, Hugh Rockoff, and Angela Redish (1994), "The U.S. Banking System from a Northern Exposure: Stability versus Efficiency," *Journal of Economic History* 54, 325–41.

Borio, Claudio and Mathias Drehmann (2009), "Assessing the Risk of Banking Crises—Revisited," *BIS Quarterly Review,* March, 29–46.

Boyd, John, Gianni De Nicolò, and Elena Loukoianova (2011), "Banking Crises and Crisis Dating: Theory and Evidence," International Monetary Fund, revised working paper.

Boyd, John, Sungkyu Kwak, and Bruce Smith (2005), "The Real Output Losses Associated with Modern Banking Crises," *Journal of Money, Credit and Banking* 37, 977–999.

Brunnermeier, Markus and Yuliy Sannikov (2010), "A Macroeconomic Model with a Financial Sector," Princeton University, working paper.

Brunnermeier, Markus, Thomas Eisenbach, and Yuliy Sannikov (2012), "Macroeconomics with Financial Frictions: A Survey," Princeton University, working paper.

Caballero, Ricardo J., Takeo Hoshi, and Anil K. Kashyap (2008), "Zombie Lending and Depressed Restructuring in Japan," *American Economic Review* 98, 1943–77.

Calomiris, Charles, and Gary B. Gorton (1991). "The Origins of Banking Panics: Models, Facts, and Bank Regulation," in Financial Markets and Financial Crises, ed. Glenn Hubbard. Chicago: University of Chicago Press: 93–163.

Calvo, Guillermo (1995), "Varieties of Capital-Market Crises," in G. Calvo and M. King, eds., The Debt Burden and its Consequences for Monetary Policy (St. Martins Press: New York).

Cannon, James Graham (1910), Clearing Houses (Washington, DC: Government Printing Office).

Capie, Forrest and Geoffrey Woods, editors, (2007), The Lender of Last Resort (Routledge; London and New York).

Caprio, Gerard and Daniela Klingebiel (1996), "Bank Insolvencies: Cross-Country Experience," World Bank Policy Research Working paper PRWP1620.

Caprio, Gerard and Daniela Klingebiel (1999), "Episodes of Systemic and Borderline Financial Crises," World Bank, working paper.

Carlsson, Hans and Eric van Damme (1993), "Global Games and Equilibrium Selection," *Econometrica* 61 (5), 989–1018.

Cassis, Youssef (2011), Crises and Opportunities: The Shaping of Modern Finance (Oxford University Press).

Cerra, Valarie and Sweta Saxena (2008), "Growth Dynamics: The Myth of Economic Recovery," *American Economic Review* 98, 439–57.

Claessens, Stijn, M. Ayhan Kose, and Marco Terrones (2011), "Financial Cycles: What? How? When?," International Monetary Fund Working Paper No. WP/02/20.

Collyns, Charles and Abdelhak Senhadji (2002), "Lending Booms, Real Estate Bubbles, and the Asian Crisis," International Monetary Fund Working Paper No. WP/02/20.

Dang, Tri Vi, Gary B. Gorton, and Bengt Holmström (2012), "Ignorance and the Optimality of Debt," Working paper, Yale and MIT.

Dasgupta, Amil (2004), "Financial Contagion through Capitala Connections: A Model of the Origin and Spread of Bank Panics," *Journal of the European Economic Association* 2, 1049–84.

Dell'Ariccia, Giovanni, Enrica Detragiache, Raghuram Rajan (2008), "The Real Effect of Banking Crises," *Journal of Financial Intermediation* 17, 89–112.

Demirgüç-Kunt, Asli, and Enrica Detragiache (1998), "The Determinants of Banking Crises: Evidence from Developing and Developed Countries," *IMF Staff Papers* 45 (1): 81–109.

Demirgüç-Kunt, Asli and Enrica Detrachiache (2002), "Does Deposit Insurance Increase Banking System Stability? An Empirical Investigation," *Journal of Monetary Economics* 49, 1373–406.

Demirgüç-Kunt, Asli and Enrica Detrachiache (2005), "Cross-Country Empirical Studies of Systemic Bank Distress: A Survey," *National Institute Economic Review,* No. 192, April.

DeRosa, Luigi (2001), "The Beginnings of Paper Money Circulation and Neapolitan Banks," *Journal of European Economic History* 30, 497–532.

Diamond, Douglas, and Philip Dybvig (1983), "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy* 91,401–19.

Diamond, Douglas and Raghuram Rajan (2001), "Liquidity Risk, Liquidity Creation, and Financial Fragility: A Theory of Banking," *Journal of Political Economy* 109, 287–327.

Diaz-Alejandro, Carlos (1985), "Good-Bye Financial Repression, Hello Financial Crash," *Journal of Development Economics* 19, 1–24.

Djiwandono, J. Soedradjad (2000), "Bank Indonesia and the Recent Crisis," *Bulletin of Indonesian Economic Studies*, 36, 47–72.

Dornbusch, Rudi (2001), "A Primer on Emerging Market Crises," NBER Working Paper No. 8326.

Elekdag, Selim and Yiqun Wu (2011), "Rapid Credit Growth: Boon or Boom-Bust,?" IMF Working Paper WP/11/241.

Enoch, Charles, Barbara Baldwin, Olivier Frécaut, and Arto Kovanen (2001), "Indonesia: Anatomy of a Banking Crisis—Two Years of Living Dangerously, 1997–99," IMF Working Paper No. WP/01/52.

Federer, J. Peter (2003), "Institutional Innovation and the Creation of Liquid Financial Markets: The Case of Bankers' Acceptances," *Journal of Economic History* 63, 666–94.

Flandreau, Marc and Stefano Ugolini (2011), "Where It All Began: Lending of Last Resort and the Bank of England during the Overend, Gurney Panic of 1866," The Graduate Institute, Geneva, working paper No. 04/2011.

Freixas, Xavier, Bruno Parigi, and Jen-Charles Rochet (2000), "Interbank Relations, and Liquidity Provision by the Central Bank," *Journal of Money, Credit and Banking* 32, 611–38.

Furceri, Davide and Aleksandra Zdzienicka (2009), "The Effect of Banking Crises on Human Capital," working paper.

Goldstein, Itay and Ady Pauzner (2005), "Demand-Deposit Contracts and the Probability of Bank Runs," *Journal of Finance* LX, 1293–1327.

Gorton, Gary B. (1984), "Private Bank Clearinghouses and the Origins of Central Banking," *Business Review—Federal Reserve Bank of Philadelphia,* January/February, 3–12.

Gorton, Gary B. (1985), "Clearinghouses and the Origin of Central Banking in the United States," *Journal of Economic History* 45, 277–83.

Gorton, Gary B. (1988), "Banking Panics and Business Cycles," *Oxford Economic Papers* 40 (4), 751–81.

Gorton, Gary B. (1996), "Reputation Formation in Early Bank Note Markets," *Journal of Political Economy* 104, 346–97.

Gorton, Gary B. (1999), "Pricing Free Bank Notes," *Journal of Monetary Economics* 44, 33–64.

Gorton, Gary B. (2010), Slapped by the Invisible Hand: The Panic of 2007 (New York: Oxford University Press).

Gorton, Gary B. (2012), Misunderstanding Financial Crises (New York: Oxford University Press; forthcoming 2012).

Gorton, Gary B., and Andrew Metrick (2010), "Haircuts," *Review—Federal Reserve Bank of St. Louis* 92 (6): 507–20.

Gorton, Gary B., and Andrew Metrick (2012), "Securitized Banking and the Run on Repo," *Journal of Financial Economics* 104, 425–51.

Gorton, Gary B., and Don Mullineaux (1987), "The Joint Production of Confidence: Endogenous Regulation and Nineteenth Century Commercial Bank Clearinghouses," *Journal of Money, Credit, and Banking* 19, 458–68.

Gorton, Gary B., and Guillermo Ordoñez (2012), "Collateral Crises," Yale Working Paper.

Gorton, Gary B., and George Pennacchi (1993), "Financial Intermediaries and Liquidity Creation," *Journal of Finance* 45, 49–72.

Gourinchas, Pierre-Olivier, Rodrigo Valdes, and Oscar Landerretche (2001), "Lending Booms: Latin America and the World," *Economia* 1, 47–99.

He, Zhiguo and Arvind Krishnamurthy (2012), "A Macroeconomic Framework for Quantifying Systemic Risk," Kellogg School, Northwestern, working paper.

Hellwig, Christian, Arijit Mukherji and Aleh Tsyvinski (2006), "Self-Fulfilling Currency Crises: The Role of Interest Rates," *American Economic Review* 96 (5), 1769–1787.

Hoggarth, Glenn, Ricardo Reis, and Victoria Saporta (2002), "Costs of Banking System Instability: Some Empirical Evidence," *Journal of Banking and Finance* 26, 825–55.

Holmström, Bengt (2008), "Discussion of 'The Panic of 2007,' by Gary B. Gorton," In Maintaining Stability in a Changing Financial System, Proceedings of the 2008 Jackson Hole Conference, Federal Reserve Bank of Kansas City.

Holmström, Bengt (2011), "The Nature of Liquidity Provision: When Ignorance is Bliss," Presidential Address, Econometric Society, ASSA meetings, Chicago, January 5–8, 2012.

Jacklin, Charles (1987), "Demand Deposits, Trading Restrictions, and Risk-Sharing," in Ed Prescott and Neil Wallace, editors, Contractual Arrangements for Intertemporal Trade (University of Minneapolis Press; Minneapolis, MN), 26–47.

Jorda, Oscar, Moritz Schularick, and Alan Taylor (2011), "When Credit Bites Back: Leverage, Business Cycles, and Crises," Federal Reserve Bank of San Francisco Working Paper No. 2011–27.

Kaminsky, Graciela, and Carmen Reinhart. 1999. "The Twin Crises: The Causes of Banking and Balance-of-Payments Problems." *American Economic Review* 89, 473–500.

Kannan, Prakash (2010), "Credit Conditions and Recoveries from Recessions Associated with Financial Crises," IMF Working Paper No. WP/10/83.

Kelley, Morgan, and Cormac Ó Gráda (2000), "Market Contagion: Evidence from the Panics of 1854 and 1857," *American Economic Review* 90 (5): 1110–24.

Keynes, John Maynard (1930), A Treatise on Money, Vol. 2, *The Applied Theory of Money* (London: Macmillan).

Kindleberger, Charles (1978), Manias, Panics, and Crashes: A History of Financial Crises *(Basic Books).*

Kindleberger, Charles (1993), A Financial History of Western Europe (Oxford University Press; 2nd edition).

Kiyotaki, Nobuhiro and John Moore (1997), "Credit Cycles," *Journal of Political Economy* 105, 211–48.

Laevan, Luc and Fabian Valencia (2008, 2012), "Systemic Banking Crises: A New Database," International Monetary Fund Working Paper 08/224 and WP/12/163.

Laevan, Luc and Fabian Valencia (2010), "Resolution of Banking Crises: The Good, the Bad, and the Ugly," International Monetary Fund Working Paper 10/146.

Lockhart, Oliver (1921a), "The Development of Interbank Borrowing in the National Banking System, 1869–1914," *Journal of Political Economy* 29, 138–60.

Lockhart, Oliver (1921b), "The Development of Interbank Borrowing in the National Banking System, 1869–1914: II," *Journal of Political Economy* 29, 222–40.

Maggiori, Matteo (2012), "Financial Intermediation, International Risk Sharing, and Reserve Currencies," Stern School, New York University, working paper.

Martin, Antoine, David Skie, and Ernst-Ludwig von Thadden (2010), "Repo Runs," Federal Reserve Bank of New York Staff Report 444.

Mendoza, Enrique and Marco Terrones (2008), "An Anatomy of Credit Booms: Evidence from Macro Aggregates and Micro Data," National Bureau of Economic Research Working Paper No. 14049.

Mendoza, Enrique and Marco Terrones (2011), "An Anatomy of Credit Booms and Their Demise," working paper.

Mills, A. L. (1908), "The Northwest in the Recent Financial Crisis," *Annals of the American Academy of Political and Social Science,* Vol. 31, Lessons of the Financial Crisis, 113–119.

Moen, Jon, and Ellis Tallman (2010), "Liquidity Creation Without a Lender of Last Resort: Clearing House Loan Certificates in the Banking Panic of 1907," Federal Reserve Bank of Cleveland Policy Discussion Paper 2010–10.

Morris, Stephen and Hyun Shin (2001), "Rethinking Multiple Equilibria in Macroeconomic Modeling," NBER Macroeconomics Annual 2000, vol. 15, Ben Bernanke and Kenneth Rogoff, editors (MIT Press).

Ó Gráda, Cormac, and Eugene White. 2003. "The Panics of 1854 and 1857: A View from the Emigrant Industrial Savings Bank." *Journal of Economic History* 63 (1): 213–40.

Postlewaite, Andy and Xavier Vives (1987), "Bank Runs as an Equilibrium Phenomenon," *Journal of Political Economy* 95, 485–491.

Quinn, Stephen and William Roberds (2008), "The Evolution of the Check as a Means of Payment: A Historical Survey," Federal Reserve Bank of Atlanta *Economic Review* 93, 1–28.

Quinn, Stephen and William Roberds (2012), "Responding to a Shadow Banking Crisis: The Lessons of 1763," Federal Reserve Bank of Atlanta Working Paper.

Rancière, Romain, Aaron Tornell, and Frank Westermann (2008), "Systemic Crises and Growth," *Quarterly Journal of Economics* 123, 359–406.

Ratnovski, Lev and Rocco Huang (2009), "Why Are Canadian Banks More Resilient?," IMF Working Paper No. WP/09/152.

Reinhart, Carmen and Vincent Reinhart (2010), "After the Fall," NBER Working Paper No. 16344, forthcoming in Federal Reserve Bank of Kansas City Economic Policy Symposium, *Macroeconomic Challenges: The Decade Ahead* at Jackson Hole, Wyoming, on August 26–28, 2010.

Reinhart, Carmen, and Kenneth Rogoff (2008), "Banking Crises: An Equal Opportunity Menace," NBER Working Paper 14587.

Reinhart, Carmen and Kenneth Rogoff (2009a), "The Aftermath of Financial Crises," NBER Working Paper No. 14656.

Reinhart, Carmen and Kenneth Rogoff (2009b), This Time is Different: Eight Centuries of Financial Folly (Princeton University Press).

Richardson, Gary (2006), "Correspondent Clearing and the Banking Panics of the Great Depression," National Bureau of Economic Research Working Paper No. 12716.

Rochet, Jean-Charles and Jean Tirole (1996), "Interbank Lending and Systemic Risk," *Journal of Money, Credit, and Banking* 28, 733–62.

Rochet, Jean-Charles and Xavier Vives (2004), "Coordination Failures and the Lender of Last Resort: Was Bagehot Right After All?," *Journal of the European Economic Association* 2, 1116–1147.

Schnabel, Isabel and Hyun Shin (2004), "Liquidity and Contagion: The Crisis of 1763," *Journal of the European Economic Association* 2, 929–68.

Schularick, Moritz and Alan Taylor (2009), "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles and Financial Cycles, 1870–2008," *American Economic Review,* forthcoming.

Schuler, Kurt (1992), "The World History of Free Banking: An Overview," Chapter in The Free Banking Experience, edited by Kevin Dowd.

Schwartz, Anna (2007), "Real and Pseudo-Financial Crises," Chapter 9 of The Lender of Last Resort, edited by Forrest Capie and Geoffrey Woods (Routledge; London and New York).

Shubik, Martin (1975), "On the Eight Basic Units of a Dynamic Economy Controlled by Financial Institutions," *Review of Income and Wealth* 21, 183–201.

Sprague, O. M. W. (1910), "History of Crises Under the National Banking System," National Monetary Commission, U.S. Senate, 61st Congress, 2d Session, Document No. 538 (U.S. Government Printing Office: Washington D.C.).

Timberlake, Richard (1984), "The Central Banking Role of Clearinghouse Associations," *Journal of Money, Credit and Banking* 16, 1–15.

Toujas-Bernaté, Joël and Hervé Joly (2011), "How Costly Are Debt Crises?," IMF Working Paper No. WP/11/280.

Upper, Christian (2006), "Contagion Due to Interbank Credit Exposure: What Do We Know, Why Do We Know It, and What Should We Know?," Bank for International Settlements, working paper.

Usher, Abbott Payson (1914), "The Origin of the Bill of Exchange," *Journal of Political Economy* 22, 566–76.

Young, Stanley (1910), "Enlargement of Clearing House Functions," *Annals of the American Academy of political and Social Science* Vol. 36, No. 3, Banking Problems, 129–34.

Ziebarth, Nicolas (2011), "The Local Effects of Bank Failures on the Real Economy: Evidence from Mississippi during the Great Depression," Northwestern, working paper.

INDEX