# Financial Intermediation and Business Cycles

by

Michael Scott Gibson

A.B., Stanford University (1988)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1993



Author ..........................................................................
Department of Economics
October 5, 1992

Certified by ..........................................................................
Olivier J. Blanchard
Professor of Economics
Thesis Supervisor

Certified by ..........................................................................
Roland Bénabou
Assistant Professor of Economics
Thesis Supervisor

Accepted by ..........................................................................
Peter Temin
Chairman, Department of Economics

# Financial Intermediation and Business Cycles

by

Michael Scott Gibson

Submitted to the Department of Economics
on October 5, 1992, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

## Abstract

The thesis contains three chapters, each of which bears on the relationship of financial intermediation to the macroeconomy. In the first chapter, I develop a dynamic model of bank behavior motivated by the relationship between a bank and a loan customer. Asymmetric information between banks and firms produces a long-term relationship. In the model, the bank faces uncertainty in loan demand and in availability of loanable funds; this uncertainty is presumed to be driven by events in the macroeconomy.

The bank maximizes profit by choosing when to add and drop customers. Because of the assumption of asymmetric information, the bank always prefers to lend to an old customer over a new customer. The bank's response to shocks to loan demand and loanable funds is nonlinear; the nonlinearity implies that banks dampen economic fluctuations. Evidence on credit availability for new and established customers and on the response of bank lending to monetary policy shocks tend to support the model's conclusions.

The second chapter also focuses on the long-term relationship between a bank and a loan customer, but from the firm's point of view. Once a bank acquires information about a firm by lending it money, the bank will be able to provide future loans at a lower cost than an outside bank. A long-term relationship between a bank and a firm takes advantage of that lower cost. The primary focus of this chapter is how the contracts between the bank and the firm will be written and how these contracts affect capital accumulation.

I embed the contracting decision of the bank and the firm in a three-period overlapping generations model. Firms borrow to finance investment, and the terms of the loan contract determine which firms can invest. The model in this chapter establishes that long-term banking relationships are important, but their importance varies across firms. Very productive firms can bypass banks; middle-productivity firms benefit the most from the lower borrowing cost a long-term relationship provides. This accords well with real-world financial arrangements.

The third chapter examines a particular historical episode where financial intermediation and the macroeconomy both suffered tremendous dislocation — the Great Depression. I test whether the Depression in the United States was worsened by the banking collapse. Previous research on this question has not adequately dealt with

the problem of simultaneity between bank failures and economic activity, which this chapter carefully does.

I test for reverse causality in two ways. The first test exploits the difference in regulation of bank branching across states to derive a test for the importance of the causality from economic activity to bank failures. The second test uses bank capital as an instrumental variable for bank failures. Bank capital is a good instrument; it should be unaffected by a contemporaneous shock to economic activity, and more strongly capitalized banks should be less susceptible to failure.

Neither test can reject the null hypothesis that reverse causality from economic activity to bank failures does not bias the coefficients of the ordinary least squares regression. This supports the conclusion of Bernanke (1983) in favor of an independent role for bank failures in deepening the Great Depression. However, neither of the tests is very powerful. The failure to reject is due to large standard errors, not to coefficient estimates that strongly suggest the absence of reverse causation.

Thesis Supervisor: Olivier J. Blanchard
Title: Professor of Economics

Thesis Supervisor: Roland Bénabou
Title: Assistant Professor of Economics

# Acknowledgments

I would like to thank Olivier Blanchard, my primary advisor, for constant encouragement as I struggled with the chapters of this thesis. His comments and advice helped me greatly as I wrote and rewrote the three chapters. Perhaps even more influential was the example he set as a teacher and a researcher. More than any class I took at MIT, Olivier's example taught me "how to be a good economist." Whether I will be able to live up to that example is now up to me.

My second advisor, Roland Benabou, provided me with a lot of help along the way. He commented in detail on each of the three chapters and was a valuable resource for the toughest technical problems. I enjoyed my interactions with the rest of the MIT faculty. I would especially like to thank Jeff Wooldridge not only for commenting on my work but for devoting so much effort to his econometrics courses. He demonstrated that any subject, no matter how difficult, can be taught well if enough effort is put into the task. I hope I never forget that.

The people who made graduate school bearable and at times fun were my fellow students. I thank them for doing what they could to make MIT a friendly place. I put a lot of value on the friendships I have made at MIT and I hope they continue well into the future.

The most important person to thank is of course, my future wife, Colette. I could never say enough about how she has helped me, but a few ways are obvious: she provided lots of encouragement, she was understanding about the time I had to devote to my work, and she lent me her talents as a proofreader and editor. Because she cheered me up when I needed it most, this thesis must be dedicated to her.

I would like to thank the National Science Foundation for providing financial support in the form of an NSF graduate fellowship.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Banking and Business Cycles: A Dynamic Model of Customer Relationships

## 1.1 A Dynamic Model of Customer Relationships

### 1.1.1 Introduction

A bank receives two loan applications: the first is from an old, familiar customer whose business the bank knows well. The second is from a firm that has never borrowed from the bank before. How will the bank choose who to lend to, and what are the macroeconomic consequences of the bank's choice? These are the questions I address in this paper. If the modern focus on asymmetric information in bank lending is right, the bank will prefer, *ceteris paribus*, to deal with the customer whose business it knows rather than invest time and effort to learn the prospective new borrower's business. The bank's preference can have implications for business cycles.

In the model I develop, a bank must pay a transaction cost to make a loan to a new customer; this represents the specific investment a loan officer must make to learn a new borrower's business. The need for a specific investment stems from firms' private information. The bank faces uncertainty in loan demand and in availability of loanable funds and chooses loan supply, with the restriction that loans never exceed total assets. The bank can control its loan portfolio by dropping a customer costlessly or by adding one and paying the transaction cost. Assets not held in loans are invested in securities and bank profits are increasing in the ratio of loans to assets. The bank maximizes profit by choosing when to add and drop customers.

The bank allows the percentage of loans in its portfolio to fluctuate within a band. Inside the band, the bank meets the loan demand of established customers but does not give credit to new customers. When the fraction of assets held as loans falls to the lower boundary of the band, the bank adds new loan customers. At the upper boundary, borrowers are dropped. It makes intuitive sense for the bank to add or drop customers infrequently, since adding customers is costly and dropping customers is unprofitable. The model posits a sharp distinction between new and established loan customers. One reason why the distinction is potentially important for the economy is that young firms who may not yet have a relationship with a bank are concentrated in high-technology, high productivity industries.

In the first half of the paper, I develop a dynamic model of bank behavior based on the customer relationship. I look at the differential treatment of new and old customers and its macroeconomic consequences. Because banks respond nonlinearly to shocks to loan demand and loanable funds, the model has implications for business cycles. In the second half of the paper, I look at some evidence on credit availability for new and established customers. I examine bank behavior around a particular set of monetary policy shocks, those identified by Romer and Romer (1989). Finally, I look at postwar U.S. data to see whether the model's implications for business cycles are borne out.

### 1.1.2 Can banks matter for business cycles?

What are the links between financial intermediation and business cycles? Bankers believe the banking sector influences business cycle fluctuations, but economists who try to model business cycles often ignore the banking sector (e.g. Prescott (1986)). Eckstein and Sinai (1986) and Wojnilower (1980) tell rich stories of business cycles caused by regulatory rigidities or collapses in important financial markets. Recent empirical work has provided support for the idea that banks can influence business cycles (Bernanke (1983)), but the evidence is not compelling on either side of the question.

Eckstein and Sinai (1986) and Wojnilower (1980) describe vividly how the financial sector contributed to post-World War II business cycles. Both assign the financial sector a key role in causing and propagating recessions. They eschew econometric methods, arguing that so many structural changes have taken place in financial markets over the postwar period that no structural parameters could be stable across the entire sample. I use similar descriptive techniques along with a simple vector autoregression to investigate my model's conclusions.

Bernanke and Gertler (1989) and Williamson (1987) both construct real business cycle models in which asymmetric information and costly state verification make financial intermediation necessary. In Bernanke and Gertler's paper, a good shock raises income, improving borrowers' balance sheets and making it easier for them to finance investment; this continues the boom. In Williamson's paper, a good shock makes investment projects less risky, enabling more entrepreneurs to get credit. In both models, technology shocks would not create business cycles without financial intermediation. In my model, the opposite result obtains. Banks dampen the business cycle most of the time when asymmetric information makes long-term relationships desirable.

## 1.2 Long-term customer relationships

Among financial intermediaries, only banks maintain long-term relationships with their loan customers. This fact makes banks special. Because establishing a long-term relationship is costly, banks prefer to lend to established customers rather than to new borrowers. Before moving on to the dynamic model of bank behavior with long-term customer relationships in section 1.3, I explore why banks and borrowers enter into long-term relationships. I also examine past research that has taken some aspect of customer relationships into account when modeling bank behavior.

### 1.2.1 Switching costs or specific investments give rise to long-term relationships

Tirole (1988, p. 21) remarks that "long-run relationships are often associated with either switching costs or specific investments." In the case of banks and their loan customers, both sources of long-run relationships stem from asymmetric information, so they cannot be easily separated. Banks cannot know a firm's profitability without actually lending to it. Once a bank decides to lend to a firm, it must spend resources to learn the firm's business. A bank acquires non-transferable knowledge, which makes it optimal for the bank and the borrower to continue their relationship for as long as possible.

Neither the bank nor the borrower will want to end the relationship once the initiation costs have been paid. If the bank cuts off the credit of an established borrower it must find a new place for those funds. To make another loan, it must find a new loan customer and spend the resources to initiate a loan. If the borrower switches its borrowing from one bank to another, it must prove its creditworthiness to the new lender.

### 1.2.2 Micro models of long-term relationships in banking

Many authors have studied how asymmetric information in bank lending gives rise to long-term relationships. For example, Haubrich (1989) and Sharpe (1990) construct models where long-term relationships are preferred to spot lending because of an asymmetry of information. Their models differ in structure but have some similar implications. In both papers, a long-term relationship is costly to establish because of the asymmetry of information between banks and firms. A long-term relationship is optimal because it minimizes those costs, giving a firm access to cheaper credit.

### 1.2.3 Previous studies of long-term relationships

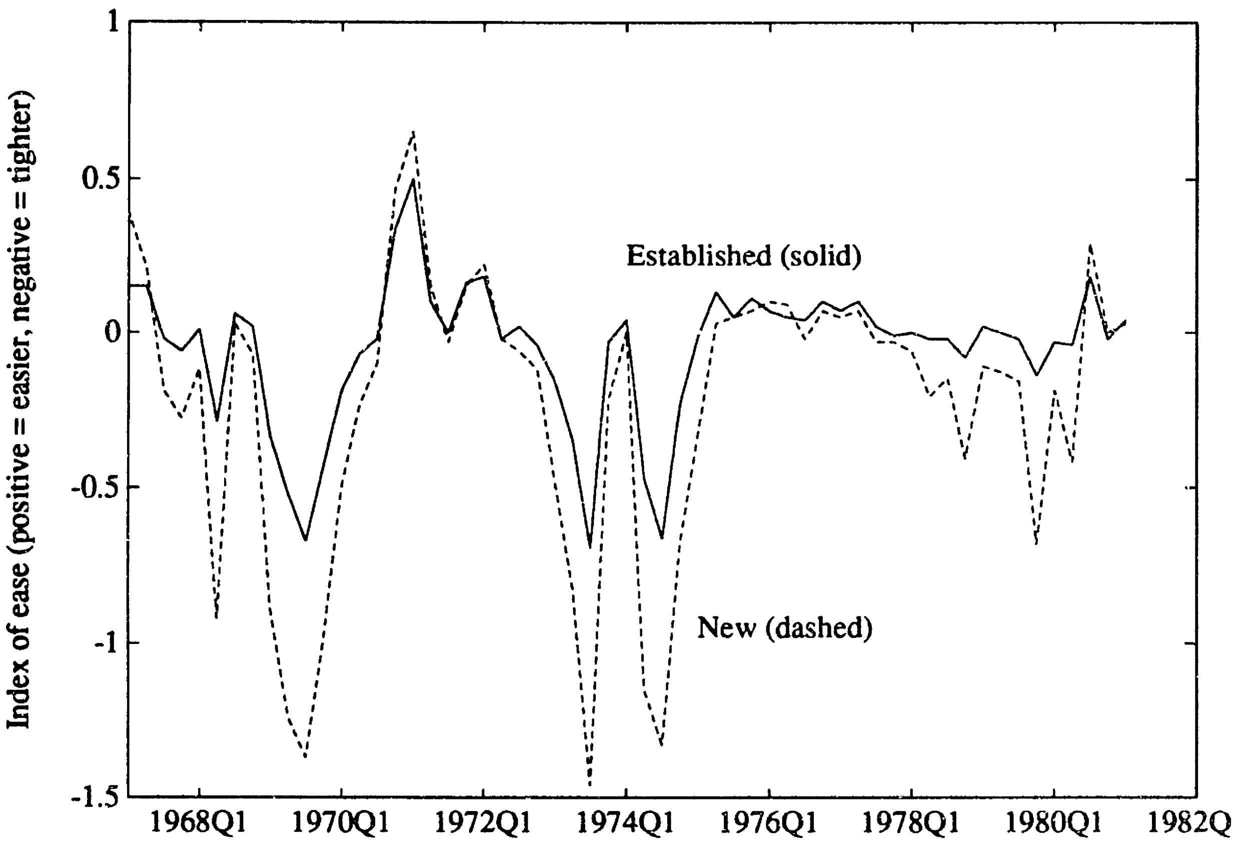
Several authors have considered various effects of long-term relationships in banking. The leading study on this subject is Hodgman (1963). He bases his discussion of bank lending policy on extensive interviews with managers at 18 large U.S. commercial banks. The striking feature of the interviews is how strongly bankers emphasize the customer relationship. Hodgman also presents a static model of bank lending behavior which accounts for customer relationships and focuses on the relationship between lending and deposit balances. Hodgman's interview findings provide a useful insight into banking practice, and I will refer to them to motivate several features of my model. Kane and Malkiel (1965) and Wood (1975) both model bank behavior, taking customer relationships into account. My model complements these earlier efforts because it is explicitly dynamic, it incorporates uncertainty in a rigorous way, and it links bank behavior motivated by customer relationships to business cycles.

### 1.2.4 Some evidence on long-term relationships

Long-term customer relationships should drive a wedge between credit availability for new loan customers and credit availability for established loan customers. Before describing my model, I present some evidence on this point in figure 1-1. For fourteen years (1967–1980), the Federal Reserve's quarterly survey of bank lending practices asked bank loan officers whether credit had become easier or tighter for new and established customers. From the survey responses, I made two time series, displayed in figure 1-1, which measure ease of credit for new and established borrowers.[1] Two results emerge from the figure. First, new borrowers consistently face tighter credit conditions than established borrowers. The dashed line lies below the solid line almost all the time. This confirms the basic premise of my model, that banks prefer to lend to established customers. Second, credit conditions for new borrowers should vary a lot more than credit conditions for established borrowers in a model with long-term relationships; this is true according to figure 1-1. The dashed line has a larger

[1] See the data appendix for details of construction of these time series.

Figure 1-1: Ease of credit for new and established borrowers



amplitude than the solid line. Keeping in mind the evidence on lending to new and established customers, I turn now to a full development of a model of banking with long-term relationships.

## 1.3 Bank behavior with long-term relationships

### 1.3.1 Definitions

The model highlights the effects of long-run banking relationships, as discussed above, on bank behavior. At time $t$, the bank has $c(t)$ loan customers, with an average loan size of $\lambda(t)$. The size of the bank's loan portfolio $c(t)\lambda(t)$ must always be less than its total assets $A(t)$. Let $z(t)$ be the ratio of loans to assets,

$$z(t) \equiv \frac{c(t)\lambda(t)}{A(t)}. \tag{1.1}$$

The bank holds all assets not lent to firms in securities, so its securities portfolio equals $A(t)[1-z(t)]$. The bank faces uncertainty in the average loan size $\lambda(t)$ and total assets $A(t)$. Let $\lambda(t)$ and $A(t)$ be geometric Brownian motions with parameters $(\mu_\lambda, \sigma_\lambda)$ and $(\mu_A, \sigma_A)$ respectively. These two variables move randomly, beyond the bank's control. The bank can only control $z(t)$ by varying the number of loan customers, $c(t)$.

#### Laws of motion

The laws of motion for $c(t)$, $\lambda(t)$ and $A(t)$ are assumed to be:

$$\begin{aligned}
\frac{dc(t)}{dt} &= -\delta c(t) \\
\frac{d\lambda(t)}{\lambda(t)} &= \mu_\lambda dt + \sigma_\lambda dw_\lambda(t) \\
\frac{dA(t)}{A(t)} &= \mu_A dt + \sigma_A dw_A(t)
\end{aligned}$$

where $w_\lambda(t)$ and $w_A(t)$ are Wiener processes with correlation $\phi$. Use Ito's lemma and equation (1.1) to calculate the law of motion of $z(t)$ when no customers are added or dropped:

$$\begin{aligned}\frac{dz(t)}{z(t)} &= \mu_z dt + \sigma_z dw_z(t), \qquad (1.2)\\ \text{where } \mu_z &= -\delta + \mu_\lambda - \mu_A - \sigma_\lambda\sigma_A\phi + \sigma_A^2\\ \sigma_z &= (\sigma_\lambda^2 + \sigma_A^2 - 2\phi\sigma_\lambda\sigma_A)^{1/2}\end{aligned}$$

and $w_z(t)$ is a Wiener process.

### Is average loan size beyond the bank's control?

Hodgman (1963), in one of the early studies of the customer relationship in banking, interviewed commercial bankers and found that banks generally meet any reasonable loan request from an existing customer to avoid losing business to a competing bank. He quotes a banker as saying, "we make loans we do not care for in order to retain a good customer relationship. . . . We are not going to let other banks become the principal banking connection for our customers for whom we are now the principal banking connection."[2] Jaffee (1971, p.10) remarked that, due to lines of credit and prenegotiated loan contracts, "in the short run the banks are obliged to meet the loan requests of these [long-term] customers." It is rare for a bank to refuse a request for a credit increase. A borrower with a loan commitment cannot be denied funds without breaking a written contract. On the other hand, if a borrower wants to pay down its outstanding loan balance, there is little the bank can do about it.

### The bank has total control of the number of loan customers

I assume the bank can increase or decrease $c(t)$ at any time. The bank can decrease $c(t)$ by simply dropping a customer. To be able to increase $c(t)$ at any time, the bank must be able to find a willing borrower in booms and in slumps. This condition will always be satisfied if credit rationing is pervasive. There is a constant separation rate of $\delta$ representing borrowers who sever their relationship with the bank.

[2] A banker, quoted in Hodgman (1963), p. 30.

### Total assets fluctuate randomly

I assume $A(t)$ is random and cannot be adjusted by the bank. The motivation for this assumption is that depositors can withdraw their funds from a bank at any time. The portfolio choices of depositors and their wealth are beyond the bank's control. My assumption rules out active liability management, where a bank lends money first and then incurs a liability (sells large CDs, borrows federal funds or Eurodollars) to keep its balance sheet balanced.[3]

### The bank's objective function

The bank's decision problem at time $t$ is

$$\max_{\{c(s):s=t,\ldots,\infty\}} E_t \left[ \int_t^{\infty} e^{-\rho s} f(z(s)) ds - \Gamma \right] \tag{1.3}$$

subject to $0 \leq z(s) \leq 1, \ \forall s$

and (1.2), the law of motion for $z(t)$,

where $\rho$ is the bank's discount rate, $f(z)$ is the bank's profit from putting a fraction $z$ of its assets into loans, and $\Gamma$ is the cumulative cost of adding customers. Assume $f' > 0$, so the bank earns more profit by making more loans. This implies loans pay a higher interest rate than securities. Think of $f(z)$ as

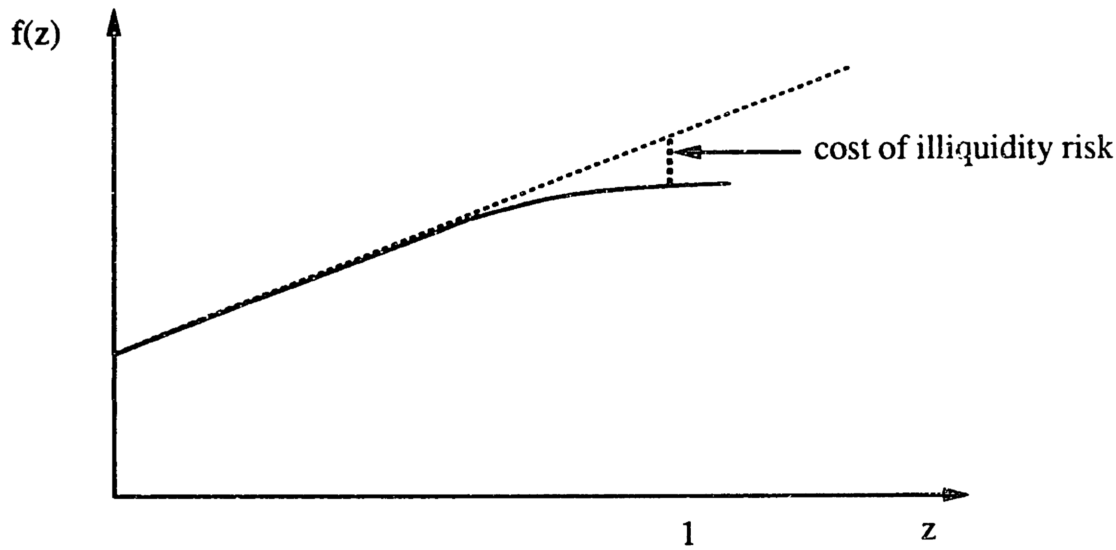
$$f(z) = i_L z + i_S(1 - z) - \psi(z),$$

where $i_L$ and $i_S$ are (net) interest rates earned on loans and securities and $\psi(z)$ is a weakly increasing, positive function representing the expected cost of illiquidity risk. As $z$ rises, the bank's assets become less liquid and the chance of costly illiquidity rises. Figure 1-2 sketches what $f(z)$ should look like.

By assuming that the profit function remains constant over time, I am implicitly assuming that the spreads between loan rates, bond rates, and cost of funds also

[3]One way to introduce liability management into the model might be to allow $\phi$, the correlation between the stochastic parts of loan demand and loanable funds, to be a choice variable.

Figure 1-2: The profit function $f(z)$




remain constant.[4] There are three reasons why omitting interest rates from the model is not a serious problem. First, theoretical models of credit rationing suggest interest rates will not move to clear credit markets. Lenders cannot raise interest rates without affecting the quality of their loan portfolio. Second, if you ask bankers how they deal with excess demand for loans (as Hodgman did), they do not say that they raise interest rates until supply equals demand. They control loan volume by tightening collateral or compensating balance requirements but rarely by changing interest rates or loan maturities. "Interest rate is not useful to control volume."[5] This second point is undoubtedly related to the first; bankers understood long before economists proved it formally that using interest rates to clear the market for loans is not optimal behavior. Third, since many loans are granted under loan commitments, which specify interest rate spreads in advance for the duration of the commitment, bankers have less scope than one might think for changing rates on short notice, at the business cycle frequency. Wojnilower (1980) agrees that interest rates take a back seat to credit availability for explaining business cycles.

[4]I am also assuming liquidity needs (the function $\psi$) and default risk are constant.

[5]A banker, quoted in Hodgman (1963), p. 30.

### 1.3.2 Add long-term relationships

I have not yet added long-term relationships between banks and their loan customers to the model. In section 1.2 I discussed the microeconomic motivations behind long-term relationships. Now I will make a crude simplification of the effects of asymmetric information on bank behavior: assume the bank must pay a cost $\gamma$ any time it adds a new loan customer. Customers can be dropped costlessly. Adding this friction to the bank's decision problem will affect bank behavior.

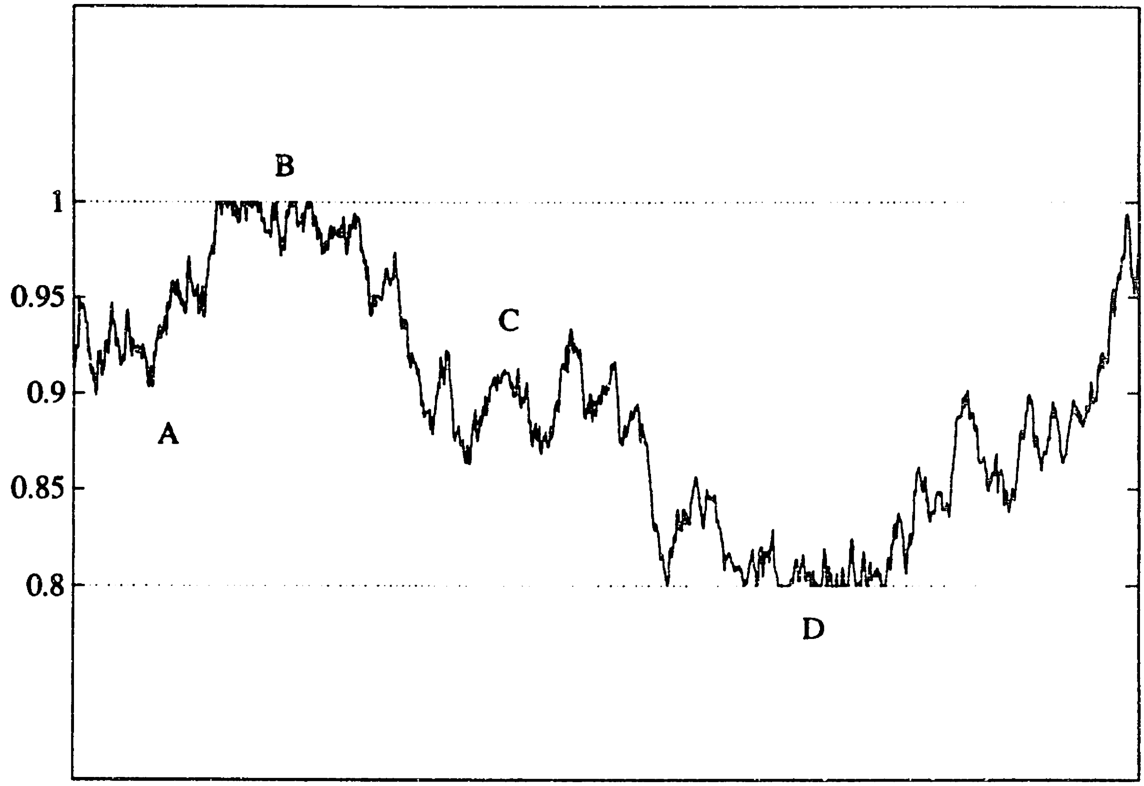
Without the cost $\gamma$ of adding a new loan customer, the bank's optimal policy would be simple: keep $z(t) = 1$ at all times (since $f' > 0$). With the cost of adjustment $\gamma$, the bank's problem falls into the class of regulated Brownian motion problems studied by Harrison (1985) and Dumas (1989). Economists have used this mathematical technique to study inventories, labor demand, entry and exit decisions and irreversible investment.[6] This paper is not intended to offer any new results in the mathematics of regulated Brownian motion but rather to apply existing results to the analysis of long-run banking relationships. Using the existing results, solving the problem (1.3) with the transaction cost $\gamma$, the bank's optimal policy is to let $z(t)$ fluctuate freely between some bounds $u$ and $l$ and to control $z(t)$ by adding or subtracting customers when it hits the bounds.[7] Figure 1-3 displays the optimal policy. (Ignore the letters A-D for now.)

The bank chooses $u$ and $l$ optimally. It is easy to see that the bank will set $u = 1$. Since the bank can costlessly drop customers at any time, it will wait as long as possible to do so, keeping $z(t)$ high and profits high in the meantime. The bank's choice of $l$ will depend on the parameters of the model; a derivation of the optimal $l$ is given in the mathematical appendix, along with comparative statics results. Most importantly, a larger $\gamma$ will cause $l$ to fall, because the bank will wait longer to add new customers if the cost of adding those customers is higher. As noted above, if $\gamma = 0$ then $l = u = 1$.

[6] See Constantinides and Richard (1978), Bentolila and Bertola (1990), Dixit (1989) and Bertola (1988).

[7] A more complete solution of (1.3) is given in the mathematical appendix.

Figure 1-3: A sample path of $z(t)$, with $u = 1$ and $l = 0.8$.




### 1.3.3 A description of optimal behavior

As $\lambda(t)$ and $A(t)$ fluctuate, moving $z(t)$ inside its band, the bank does not add or drop customers. When loan demand from established customers increases, the bank grants the credit. When total assets go up, the bank puts the increment into securities until it is able to loan it out. Within the band, the bank meets credit demands of established customers but does not grant loans to any new customers. When $z(t) = l$, the bank adds customers. Loan applications of new borrowers are approved. When $z(t) = u$, the bank must drop customers because it has no more money to loan out. I assume the relationship between a firm and a bank must be continuously maintained for the firm to remain an "insider" borrower. The bank retains no knowledge of the business of a dropped customer.[8] As in all models of this form, adding a linear transaction cost to the agent's decision problem causes the agent to adjust infrequently.

The bank transmits the shocks it receives in a very nonlinear fashion. If a positive shock to $\lambda(t)$ hits the bank, so old customers demand more credit, the bank's response will vary depending on whether $z(t)$ is at the top of its band or inside its band. Inside the band, the bank's loan portfolio $c(t)\lambda(t)$ changes one-for-one with the increased demand of old customers. At the top of the band, the bank's loan portfolio cannot expand at all when a positive shock to $\lambda(t)$ hits. Some old customers must be denied credit, but total credit is constant. If the bank receives a negative loan demand shock, it will not take any action inside the band other than reinvesting the funds in securities. At the bottom of the band, the bank will make a loan to a new customer to keep total credit from falling. When the bank receives a positive shock to loanable funds, $A(t)$, the bank will put the new funds into securities until an old customer asks for more credit if $z(t)$ is inside its band. If $z(t)$ is at the bottom of its band, the bank will loan out the new funds by making a loan to a new customer. A negative shock to $A(t)$ will similarly be absorbed in the interior of the band and passed on by contracting total credit at the top of the band. The transmission is nonlinear, depending on where $z(t)$ is within its band.

## 1.4 Cyclical behavior

We can use the model to analyze the movement of bank lending over the business cycle. If the randomness faced by the bank behaves like a business cycle, the bank will act differently in different phases of the cycle. When credit is tight, the bank will avoid the cost of lending to new borrowers, though existing loan customers may still be able to get credit. This "insider-outsider" effect is a stark prediction of the model.

[8] If banks retained knowledge of dropped customers, little would change in the model. Banks could costlessly choose $c(t) \in [0, K(t)]$, where $K(t)$ is the number of customers of whose business the bank has current or recent knowledge. $c(t)$ would only be less than $K(t)$ when $z(t) = u$. Once a negative shock hit $z(t)$, the bank would immediately add back old customers to keep $z(t) = u$; as soon as $c(t) = K(t)$ and all old customers were again being served, the usual dynamics would apply again.

All randomness, from the business cycle or other sources, enters the model through the random terms $dw_\lambda$ and $dw_A$. These terms make average loan size $\lambda(t)$ and total assets $A(t)$ vary stochastically. I assume the trend components $\mu_\lambda$ and $\mu_A$ are constant over time, so they cannot provide business cycle variation. In section 1.7.1 I will examine the implications of relaxing this assumption and letting the drift vary over time.

### 1.4.1 If customers could be added costlessly

If adding a new customer were free, the bank would respond to movements in $\lambda(t)$ and $A(t)$ by continually adjusting $c(t)$ to keep $z(t) = 1$. The instantaneous profit function $f(z)$ is increasing, and with costless adjustment the bank has no reason not to add customers to the point where $z(t) = 1$. The bank does not care that it may have to deny an established customer credit if loanable funds fall or loan demand increases in the future. There will be no cyclical movement in $z(t)$ in the frictionless case. The bank passes on every shock to $\lambda(t)$ or $A(t)$ to the real economy by varying the number of customers who get credit.

### 1.4.2 If it is costly to add customers

When $\gamma > 0$, the bank lets $z(t)$ fluctuate freely, following equation (1.2) above, and controls $z(t)$ when it hits a boundary of the band $(l, u)$. One possible path $z(t)$ could take is shown in figure 1-3. The process $z(t)$ shown in figure 1-3 has four distinct phases, labeled A–D on the figure. During phase A, $z(t)$ is generally increasing without being controlled. Either average loan size to established customers is increasing ($\lambda(t) \nearrow$) or loanable funds are decreasing ($A(t) \searrow$). The ratio of loans to assets rises but no new customers are added. The bank willingly supplies existing customers with more credit but refuses to pay the cost of starting long-run relationships with new customers. In phase B, average loan size relative to loanable funds still rises but the bank can no longer supply more credit. When $z(t) = 1$ and $\lambda(t)$ increases or $A(t)$ falls, the bank must refuse to serve some existing customers. This is an extreme "credit crunch." Neither existing nor prospective loan customers can get loans, because the bank has no funds available.

In phase C, loanable funds grow ($A(t) \nearrow$) or loan demand from existing customers is weak ($\lambda(t) \searrow$), creating excess loanable funds. However, the bank is unwilling to lend these funds to a new customer because it does not yet pay to incur the cost $\gamma$ required to take on a new customer. The credit crunch continues, because many "outsider" firms that would like to get credit are denied it. "Insider" firms who ask for more credit are able to get it. In phase D, loan demand has weakened and/or loanable funds have increased sufficiently that the bank lends to new customers to keep its loan portfolio $z(t)$ at the lower bound $l$. Loanable funds are abundant and/or demand for credit is weak, so credit supply is ample.

A striking implication of the model is the "insider-outsider" effect. This result is unique to a model of banking that includes long-run lending relationships. In two of the four phases, insiders can get credit when outsiders cannot. This yields a more subtle analysis of the phenomenon of "credit crunch" which so often accompanies recessions. There are really two "flavors" of crunch: when insiders can still get credit but outsiders cannot, and when neither insiders nor outsiders can get credit. This differential effect may have implications for the business cycle fluctuations of the real economy.

A second implication is that when $z(t)$ is not at one of its boundaries, the bank responds passively to shocks. Loan demand from established customers is accomodated, and shocks to loanable funds are not passed on to borrowers. In an economy where $\gamma = 0$, all shocks that hit the bank feed right through to borrowers. The dampening effect of banks in my model contrasts with the models of Bernanke and Gertler (1989) and Williamson (1987). In those papers asymmetric information in credit markets increases the amplitude of business cycles. Taking account of long-term relationships leads to the opposite conclusion.

### 1.4.3 Business cycle randomness

Before I identify phases A-D on figure 1-3 with particular stages of the business cycle, I must discuss how the exogenous processes $\lambda(t)$ and $A(t)$ might be expected to vary over the business cycle. Average loan size $\lambda(t)$ will vary as customers borrow more or less money from the bank. Since the bank supplies credit elastically to established borrowers when $z(t)$ is inside its band, $\lambda(t)$ will vary with the demand for credit. The demand for credit is procyclical, because credit is an input into the production of goods and services, and firms will demand more credit when current and future output is expected to be high; $\lambda(t)$ is procyclical.

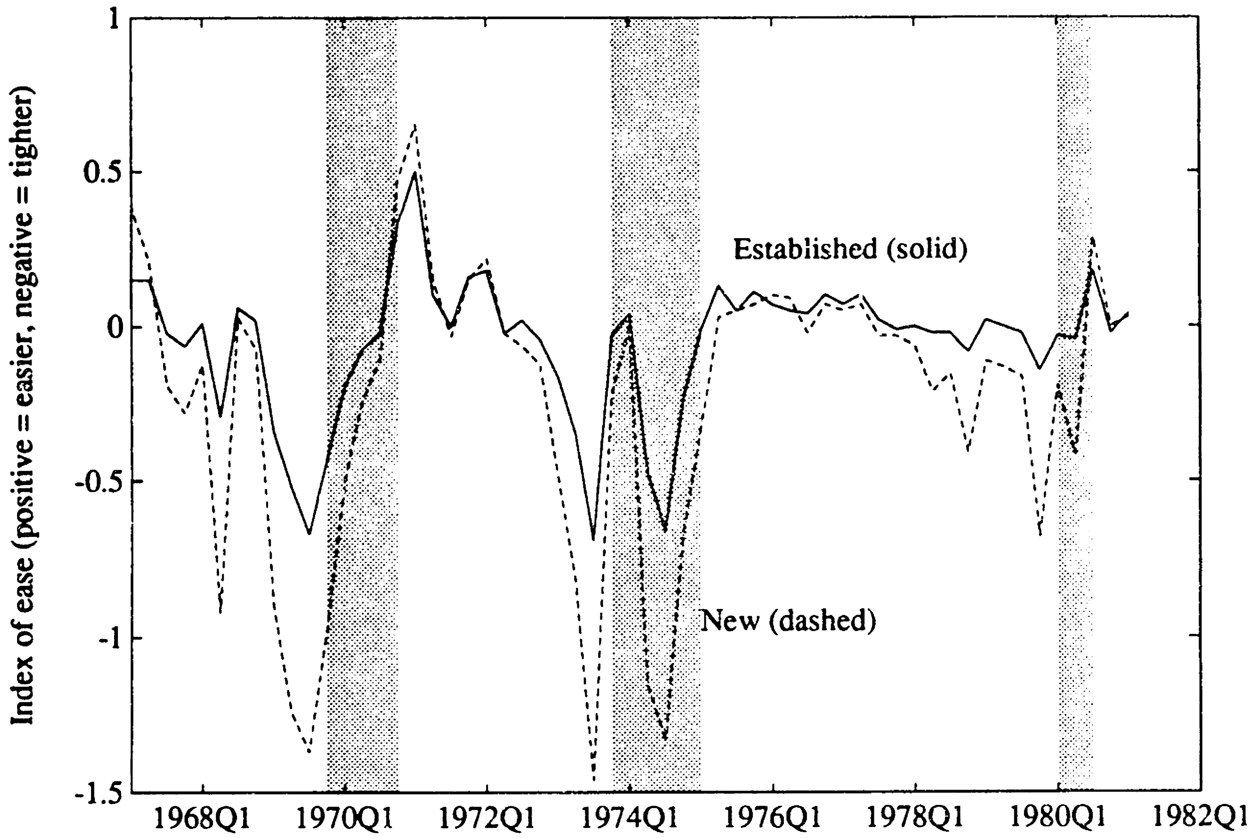
Loanable funds come from several sources for several reasons. Some come from households, some from businesses. Both businesses and households keep balances in banks for two main reasons: transactions services and savings. Both should be procyclical, transactions balances more strongly so. Another reason for $A(t)$ to vary is the compensating balances many borrowers are required to keep on deposit when they take out a loan. Since lending is procyclical, compensating balances are another reason why $A(t)$ should be procyclical.

Since both $\lambda(t)$ and $A(t)$ are procyclical and $z(t)$ depends on their quotient, nothing can be said yet about the pro- or counter-cyclicality of $z(t)$. We would have to know the magnitudes of the cycles in $\lambda(t)$ and $A(t)$, or alternatively, their income elasticities. Another way to determine the cyclicality of $z(t)$ is to exploit the bank survey evidence from figure 1-1. Figure 1-4 reproduces figure 1-1 with recessions shaded. As the figure makes clear, credit is tight for both new and existing customers during or just before recessions. In the model, credit will be tight for all customers when $z(t)$ is at the top of its band. So, recessions will occur at times when, in the model, $z(t)$ is high. This is phase B on figure 1-3.

### 1.4.4 Feedback from bank behavior to the business cycle

So far, the bank has responded to movements in loanable funds and insider loan demand by varying its asset portfolio $z(t)$. As I discussed in my introduction, some

Figure 1-4: Ease of credit for new and established customers




economists believe banks not only respond to business cycle fluctuations but contribute to them as well. Banks could contribute to booms and recessions in my model if it were extended to include feedback from bank credit to the business cycle. A bank will have an effect only when it takes action to keep $z(t)$ from crossing one of the boundaries of its band.

Consider the effect of the bank's actions on real economic activity when credit is an input into production of goods and services.[9] There are two elements to consider: which boundary $z(t)$ is at and whether the driving shock comes through $\lambda(t)$ or $A(t)$. Considering all combinations yields four cases. First, when $z = l$, new customers

[9]Blinder (1987) contains a formal model of an economy where firms require bank credit for working capital one period in advance of production.

get credit. If a positive shock to $A(t)$ hits, the loans to new customers add to total lending, so it is easy to see why output rises. Second, if $z = l$ and a negative shock to $\lambda(t)$ hits, new customers get credit because loan demand from established customers has fallen off. If $A(t)$ doesn't change, the same amount of bank lending is being distributed differently among firms. We have good reasons to believe that new borrowers who could never get bank credit, primarily young small firms, are more productive than old borrowers. New technologies are often introduced by small start-up firms. Personal computers and biotechnology, for example, are fields where young firms innovated and older larger firms struggled to keep up. When banks grant new borrowers credit (when $z = l$), they may be performing a valuable social function by providing funds to very productive firms.

If this story is correct, when $z$ hits $l$ the economy should experience a spurt of growth as new innovative firms get bank credit and expand output. As I will document in section 1.6, $z$ does not usually hit $l$ until (on average) six months into a recovery. To check whether the proposition that growth increases when $z$ reaches its lower bound is plausible, I calculated the growth rates of industrial production in the months between the business cycle trough and the trough of $z$ and in the six months following the trough of $z$. The monthly growth rate of industrial production averages .87% before $z$ hits its trough and 1.06% after. The increase is .19% per month, or 2.1% annually, which is consistent with the above analysis.[10]

The remaining two cases occur when $z = u$ and old customers are cut off from credit. If the shock that pushed $z$ up to $u$ came from $A(t)$ falling, total bank credit must fall too, so a fall in output is an obvious consequence. When $z = u$ and $\lambda(t)$ increases, some old borrowers are cut off and their credit given to other old borrowers. To see how this redistribution could have a negative effect on output, think of a simple inventory cycle model with the following characteristics. All firms use bank credit to finance inventories, but only some firms have loan commitments that guarantee their credit will never be cut off. Firms accumulate inventories if production exceeds demand, which is random. Each firm must buy intermediate

[10]The difference is statistically significant.

inputs from other firms. There are fixed costs associated with positive production, so a firm ceases production when its inventories exceed a certain threshold. After inventories have fallen sufficiently, the firm resumes production. A recession in this model is a period when demand is slack and inventories accumulate, causing firms to halt production. Recessions can occur even when credit is unlimited, *i.e.* $A(t)$ is sufficiently large that $z$ never reaches $u$.

Now let credit be provided by a bank that behaves as my model predicts and whose $A(t)$ is not large enough to keep $z < u$. If demand is weak, inventories accumulate and firms demand more credit to finance them. The bank grants the credit freely, since all firms are established customers, until $z = u$, when the bank cuts an old customer off from credit. This customer goes bankrupt and ceases production. The remaining firms face a lower demand for intermediate inputs, so their inventories accumulate *faster* than in the case of unlimited credit. Inventories are more likely to reach the threshold at which a firm closes down, so a recession is more likely. For some realizations of demand, a recession will occur in the second case but not in the first.

Many observers of the business cycle consider tightening of bank lending ("credit crunches") to play a key role in triggering recessions.[11] In the basic version of my model, banks merely respond to exogenous shocks but do not affect the business cycle. With feedback from bank lending to output added, the nonlinear response by banks to certain shocks could make a recession more likely.

### 1.4.5 Bank lending and monetary policy transmission

My model of bank behavior, combined with feedback effects of bank lending, can clarify some issues in the debate over monetary policy transmission. The debate whether monetary policy is transmitted through money or through credit has a long history; Bernanke and Blinder (1990) and Romer and Romer (1990) are two recent contributions. When examining evidence on the credit channel, both Bernanke and

[11] Eckstein and Sinai (1986), Wojnilower (1980).

Blinder and Romer and Romer assume contractionary monetary policy will force banks to cut back on lending, regardless of the present state of the economy.

My model suggests that monetary policy transmission will differ depending on where a bank's $z(t)$ lies within its band. If the money supply contracts, causing $A(t)$ to fall, $z(t)$ will rise. If $z(t)$ was already near the top of its band, the decline in $A(t)$ could push $z(t)$ up against its ceiling and cause the bank to contract lending, which will feed through the channels discussed above and reduce output. If $z(t)$ were far from the top of its band, the fall in $A(t)$ would shift $z(t)$ up but total loans would remain constant. Real output, which depends on the quantity of credit and not on $z(t)$, would be unaffected. Two episodes of equally contractionary monetary policy could have different outcomes if $z(t)$ were initially high in one episode and initially low in another.[12] Both Romer and Romer (1990) and Bernanke and Blinder (1990) estimate autoregressions that lump all periods of contractionary monetary policy together, missing this effect. I will discuss further evidence on monetary policy transmission in my model in section 1.6.3.

To try to see whether the effects of $z(t)$ on output are important, I want to carefully study movements in $z(t)$. First, however, I must take the model, which characterizes the behavior of a single bank, and see what it implies for the behavior of the aggregate data I look at in section 1.6.

## 1.5 Aggregation

The model I have just outlined describes the behavior of a single bank. Since I have no data to test the model's implications for a single bank, I will have to look at the ratio of loans to assets of all commercial banks. A similar question of aggregation was considered by Bertola and Caballero (1990) and I apply their methodology to my model in section B of the appendix. There, I show that if all banks receive a single, common shock their $z(t)$'s will be identical in the long run. In that case the dynamics of the aggregate $z(t)$, defined as the asset-weighted average of individual $z(t)$'s, will

[12] Blinder (1987) makes a similar point.

be the same as the dynamics of an individual bank's $z(t)$. If each bank receives an idiosyncratic shock, the individual $z(t)$'s will be spread out on the interval $[l, u]$ and the aggregate $z(t)$ will be constant.

To decide which of the two types of uncertainty is empirically more important, consider their sources. Aggregate uncertainty in $z(t)$ will come from national movements in average loan size and loanable funds, which will be linked to the business cycle as I discussed in section 1.4.3. Local uncertainty will come from bank-specific shocks.[13] In almost all cases, I would argue, business cycle movements will dominate local shocks. A recession is generally a nationwide event that dominates economic activity. A second way to choose between the two types of uncertainty is to look at the actual behavior of $z(t)$. If local uncertainty dominated business cycle uncertainty, the empirical cross section distribution of banks would be stable over the business cycle and $z(t)$ would be roughly constant. If local uncertainty is unimportant relative to business cycle uncertainty, $z(t)$ should move quite a bit with the business cycle. I argue, looking ahead to figures 1-6 and 1-10, that business cycle uncertainty and aggregate shocks must be very important for banks since $z(t)$ varies significantly with the business cycle. In what follows, I assume only aggregate uncertainty is present, so the dynamics of a single bank's $z(t)$ derived in section 1.3 can be applied directly to the aggregate $z(t)$.

## 1.6 Empirical tests of the model's conclusions

In this section I examine three empirical implications of the model. First, new customers will only get credit when $z(t)$ is at the bottom of its band, and some old customers will be denied credit when $z(t)$ is at the top of its band. To test this I look at data on loan commitments, which is one way to discriminate between the borrowing of new and old customers. Second, banks will transmit shocks differently depending on where $z(t)$ lies in its band. To test this I look at the effects of some

[13]Banks can also be affected by regional shocks; to model this the correlation between bank $i$ and bank $j$ would be a function of the distance between the two. For simplicity I omit regional shocks.

specific shocks, the monetary policy shocks identified by Romer and Romer (1989). Third, the relative timing of movements in $z(t)$ and output at the business cycle frequency can tell us whether the mechanism for banks to affect business cycles outlined in this paper could have been operative in the postwar U.S. economy. Before discussing these implications, I introduce the time series for $z(t)$.

### 1.6.1 Data on $z(t)$

I construct a time series for $z(t)$ using data from the Federal Reserve.[14] [15] The individual data series were seasonally adjusted by the Census Bureau's X-11 procedure,[16] and a time series for $z(t)$ was constructed as:

$$z(t) = \frac{\text{Loans}}{\text{Loans} + \text{U.S. Government Securities} + \text{Other Securities}}.$$

The behavior of $z(t)$ over the last 130 years is presented in figures 1-5 and 1-6.

The fraction of earning assets held in loans, $z(t)$, has varied significantly since the Civil War, rising from 0.45 in 1863 to over 0.85 in the late nineteenth century and falling dramatically during the Great Depression. In the postwar period, $z(t)$ has had an upward trend of about 0.01 per year. At the end of World War II, banks' portfolios were unnaturally imbalanced, as they were heavily weighted with government debt. Since 1948, $z(t)$ has slowly returned to historic levels. The ratio seems to have stabilized in the past five years at about 0.77.

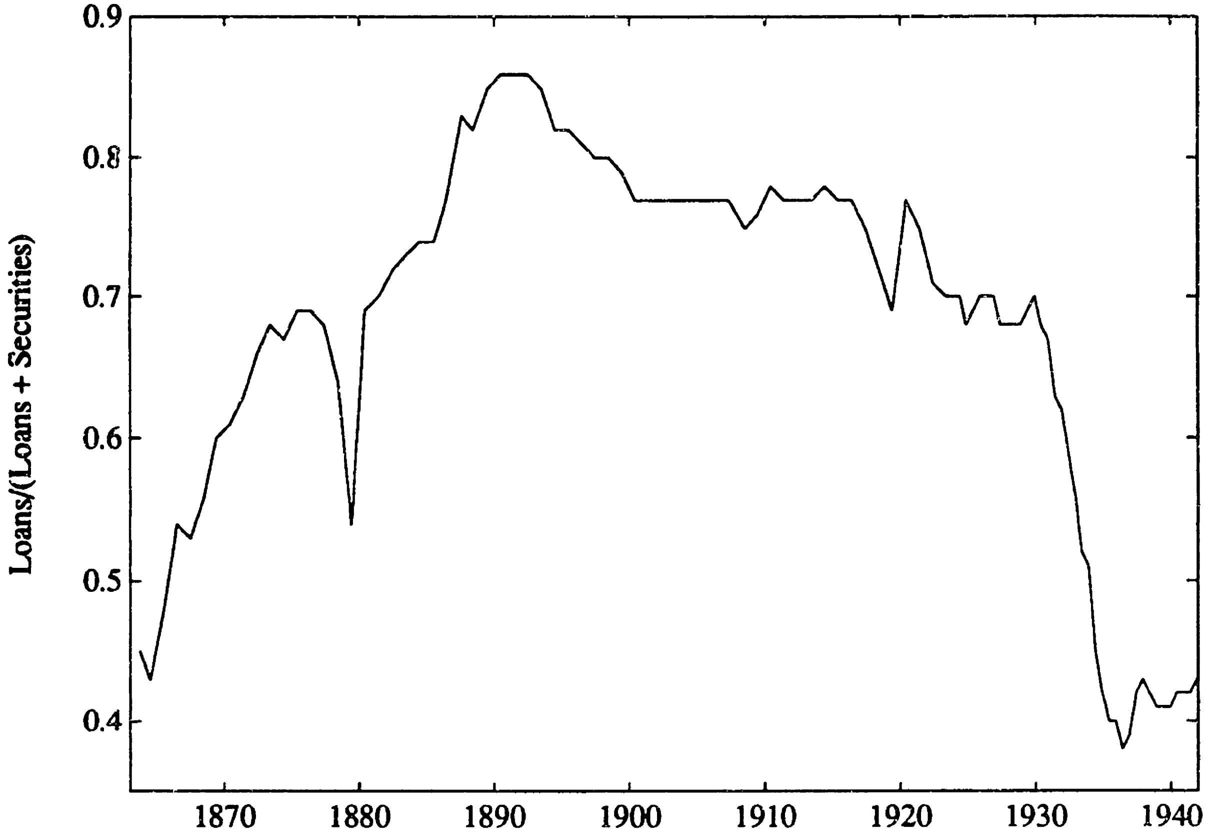
These long swings in $z(t)$ over decades are not what my model tries to explain. A likely explanation for why $z$ is always far below one is that regulation or behavioral inertia produces an upper bound on $z(t)$, $\hat{u}(t)$, that varies through time as bankers' attitudes and regulatory environments evolve. The presence of $\hat{u} < 1$ would change the constraint in the bank's optimization problem to $0 \leq z(t) \leq \hat{u}(t)$ but would not

[14]Data sources are provided in the data appendix.

[15]I test the model's implications with macro data because micro data on bank lending that differentiates between old and new customers is hard to come by.

[16]Seasonal adjustment is desirable in this case because any seasonal factors that affect $z(t)$, such as liquidity demands related to harvests or holidays, are outside the scope of the model. As it turned out, $z(t)$ was hardly affected by the seasonal adjustment.

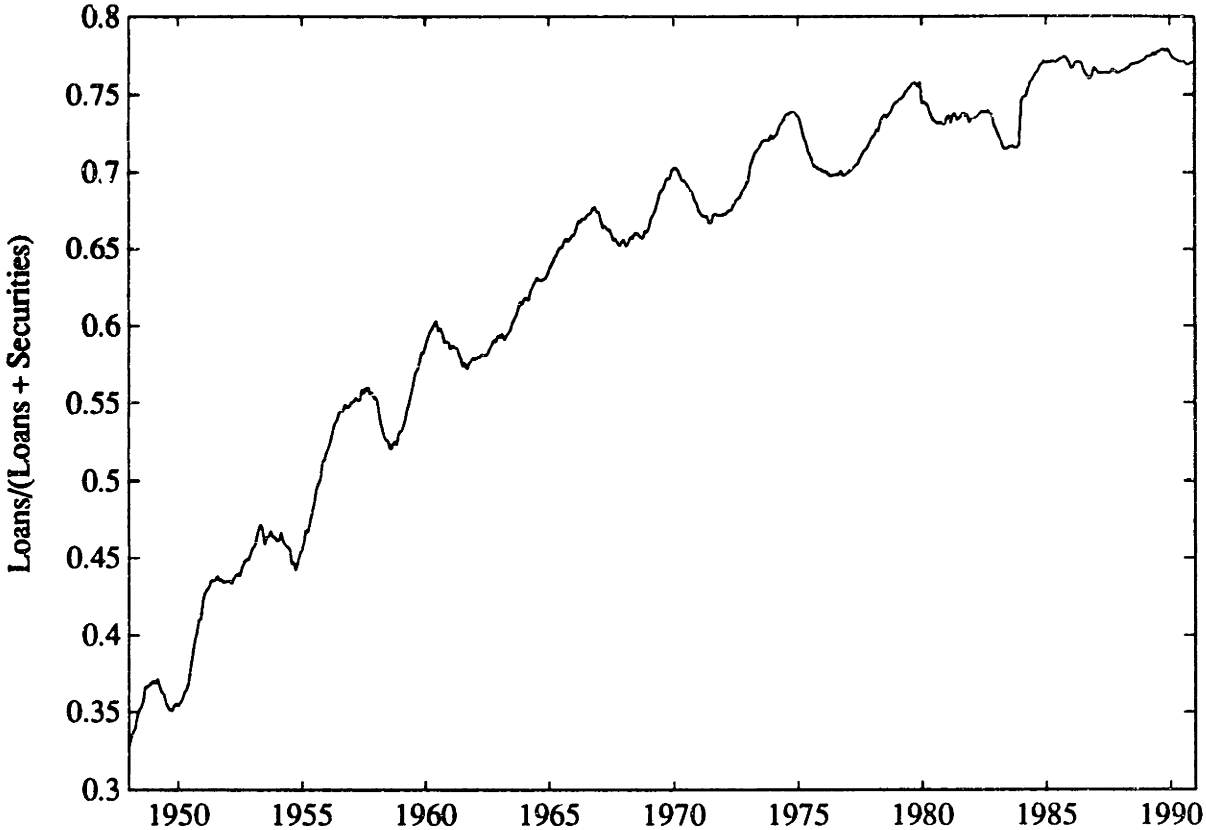
Figure 1-5: $z(t)$ for national banks, 1863-1941




otherwise affect the solution ($z$ would still be regulated in the band $(l, \hat{u})$). Apparently, $\hat{u}$ fell during the Great Depression and has risen since World War II as regulation has eased and bankers have adopted more competitive attitudes. Because of the trend in figure 1-6, I will detrend $z(t)$ when I look at cyclical movements.

A simple implication of long-run relationships in banking is that $z(t)$ should vary systematically over the business cycle, as discussed in section 1.4.3. If adding new customers were costless, $z(t)$ would be kept as close to $u$ as possible to maximize profits, so the fact that $z(t)$ varies substantially over the business cycle is itself evidence in favor of long-run banking relationships. Of course, there are other reasons why $z(t)$ could vary over the cycle which are absent from the model. If the riskiness of loans varies over the business cycle, the instantaneous profit function $f(\cdot)$ would vary

Figure 1-6: $z(t)$ for all commercial banks, 1948-1990




over the cycle as well. If loans become more risky in a recession, the optimal $z(t)$ will rise during a boom and fall during a recession. If liquidity needs rise during recessions, banks will want to hold more securities and the optimal $z(t)$ will fall during a recession.

### 1.6.2 Loan commitments

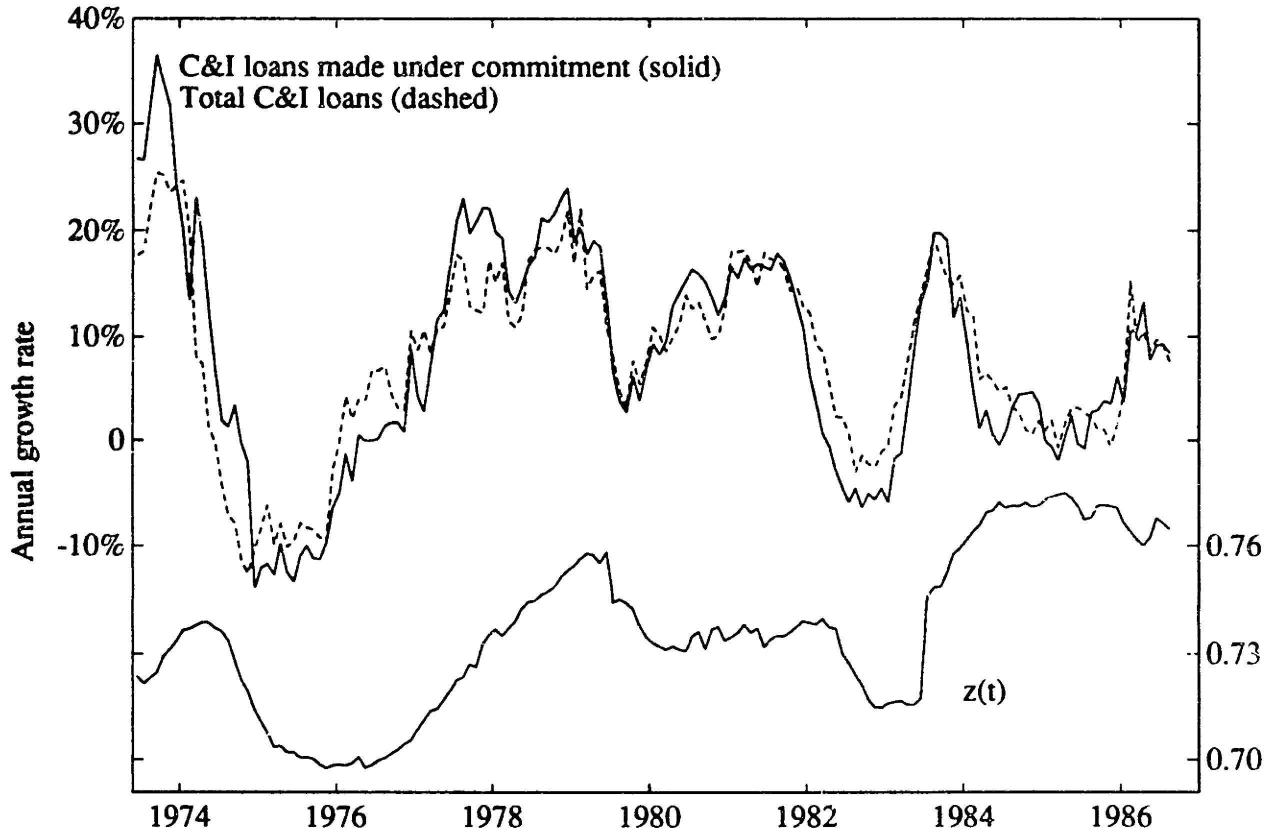
The most direct implication of the model is the difference between treatment of new borrowers and established borrowers. One type of evidence that might reveal differences between new and established borrowers is data on loan commitments. A loan commitment is a formal or informal agreement between a bank and a loan customer; the bank pre-approves the customer for credit up to a specified limit. For my purposes, the relevant point is that loans made under a commitment are loans to established customers. The cost of learning the customer's business has already been paid when a commitment is made, so a subsequent loan can be made costlessly. By comparing the growth rate in loans made under commitment to the growth rate of total loans, I can compare (net) lending to established customers with (net) lending to all customers. If the model in this paper is correct, differences in the growth rates should be systematically related to the ratio of loans to assets, $z(t)$.[17]

Loan commitment data broadly confirms the predictions of my model. Figure 1-7 displays the growth rates of commercial and industrial (C&I) loans made under commitment and total C&I loans, along with $z(t)$.[18] The tick marks at the base of the figure mark the midpoint of the indicated year. The model claims that when $z(t)$ is at a low point, new customers can get credit. Furthermore, old customers may not have to use their loan commitments to obtain a loan, so loans under commitment should grow slowly relative to total loans when $z(t)$ is low. $z(t)$ reaches a low point twice on figure 1-7: 1976 and mid-1983. At both times, the growth rate of loans made under commitments is much less than the growth rate of total loans, as expected. The model also claims that when $z(t)$ is high, old customers will have a hard time getting credit. Increased loan demand from old customers with loan commitments can divert funds away from other old customers. In this case loans under commitment should grow faster than total loans. $z(t)$ has three peaks in figure 1-7: late 1974, late 1979, and 1985–87. The first two cases fit the pattern I just described. Loans made under commitment grew faster than total loans. The third case does not fit my pattern as well because growth of loans under commitment was slower than total loans for most of that period.

[17]One way to interpret studies which have used loan commitment data to analyze the channels of monetary policy transmission, such as Sofianos, Wachtel, and Melnik (1990) and Romer and Romer (1990), is that they implicitly had in mind a model such as mine, where banks should treat new and old customers differently.

[18]Details of the loan commitment data are provided in the data appendix. Because the series for loan commitments and total loans have lots of high-frequency variation, figure 1-7 presents nine-month centered moving averages. Only C&I loans are included in the figure because data on loan commitments was only available for C&I loans.

Figure 1-7: Loan commitments and $z(t)$



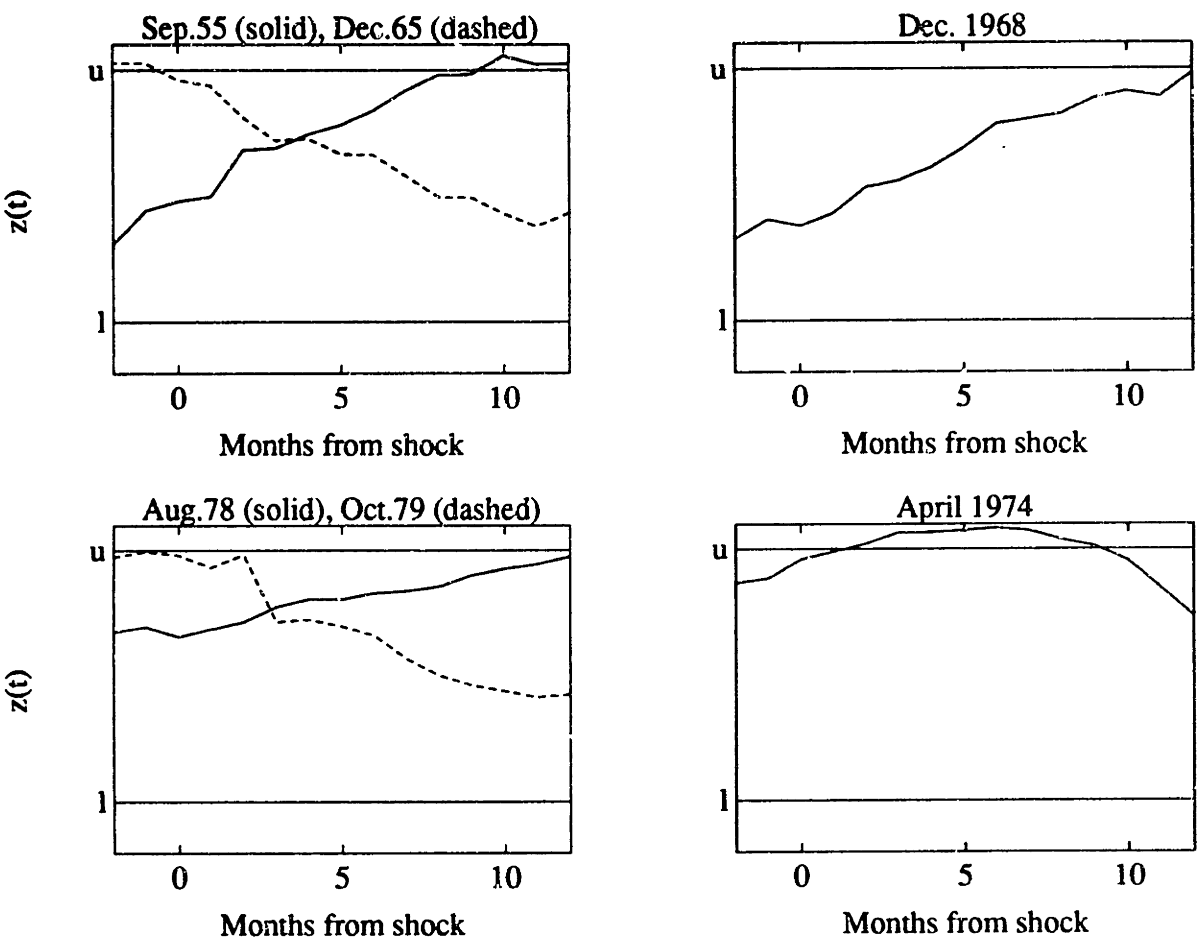

### 1.6.3 The response to monetary policy shocks

To test the model's predictions about the bank's reaction to shocks, we must be able to identify the shocks. The underlying shocks in the model, which come from loan demand $\lambda(t)$ and loanable funds $A(t)$, are hard to identify. One particular type of shock, however, can be identified. Romer and Romer (1989, 1990) have identified six episodes of contractionary monetary policy that can be taken as exogenous negative shocks to $A(t)$.[19] The path of detrended $z(t)$ from two months before each shock to twelve months after is shown in figure 1-8. The bounds $u$ and $l$ are unobserved, so I

[19] I exclude the October 1947 shock because my data begin in January 1948. I include the December 1965 shock (see Romer and Romer (1990, p. 161)).

approximated them by taking the average of the high and low points of the detrended $z(t)$ in figure 1-10.[20]

Figure 1-8: Bank behavior around the Romer dates




I have already argued that a negative shock to $A(t)$ should have a different effect on $z(t)$ (and on output) depending on where $z(t)$ is in its band. The transmission of monetary policy through credit will differ depending on where $z(t)$ is in its band. Figure 1-8 validates this hypothesis. After three episodes of contractionary monetary policy, in September 1955, December 1968, and August 1978, $z(t)$ kept rising after the shock. In all three cases $z(t)$ was well inside its band when the Fed contracted monetary policy. Clearly the shock to $A(t)$ did not cause $z$ to hit $u$; banks dampened

[20]Because $u$ has been approximated, $z(t)$ appears to go above $u$ in several cases.

the shocks. In April 1974 $z(t)$ was very close to its upper bound when the shock hit. $z(t)$ stayed at $u$ for over six months after the shock, then began to fall. Twice, in December 1965 and October 1979, $z$ was already at $u$ when the shock hit. In these episodes the model predicts that the bank cannot absorb the shock; it must pass it through into the real economy. And in both cases, $z(t)$ starts to fall shortly after the shock, exactly as we would expect if the monetary contraction had reached the economy and caused loan demand to turn down. Figure 1-8 confirms my model's conclusion that the effect of a shock is not independent of the position of $z(t)$ in its band at the time of the shock.

### 1.6.4 Cyclical movements in $z(t)$

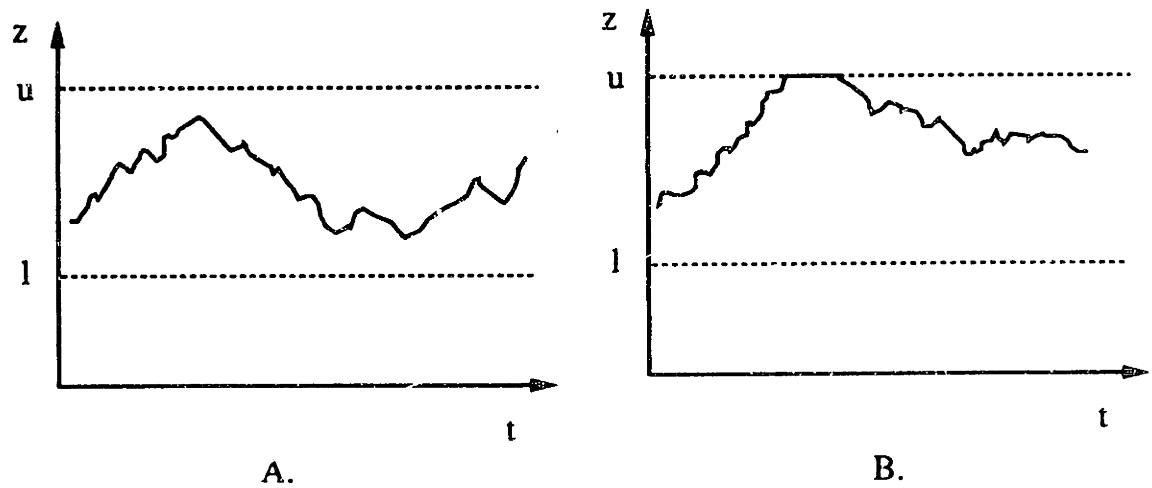
In this subsection I consider the empirical relationship of $z(t)$ with the business cycle. The purpose of this subsection is to use the model as a tool for interpreting postwar U.S. business cycles. In section 1.4.4, I discussed how banks would respond to shocks in a nonlinear fashion when $z$ hits its upper bound and cause a business cycle downturn by cutting off credit to established customers. If $z$ does not hit its upper bound but merely turns down because of shocks to loanable funds or loan demand, "insider" borrowers can consistently get the credit they request and the banking sector does not affect the business cycle. The two possibilities are illustrated in Figure 1-9.

It is hard to tell the difference between figure 1-9A and figure 1-9B by looking at a time series for $z(t)$ when $u$ and $l$ are not observed, especially if $z(t)$ is constrained above not by 1 but by $\hat{u}$, which changes over time. Figure 1-10 shows a detrended time series for $z(t)$ with recessions shaded.[21] Simply looking at figure 1-10 suggests that both situations described in figure 1-9 have occurred in the postwar era. The peak in $z(t)$ comes before the business cycle peak in 1953 and 1980.[22] These are two recessions where banks could have affected the business cycle in the way my model suggests. The peak in $z(t)$ clearly comes after the business cycle peak in 1960, 1969 and 1973.

[21]To detrend I took the residuals from a regression of $z(t)$ on time, time$^2$, and time$^3$.

[22]Whether the peak of $z(t)$ in 1948 is before or after the business cycle peak is sensitive to the detrending procedure.

Figure 1-9: Possible paths for $z(t)$

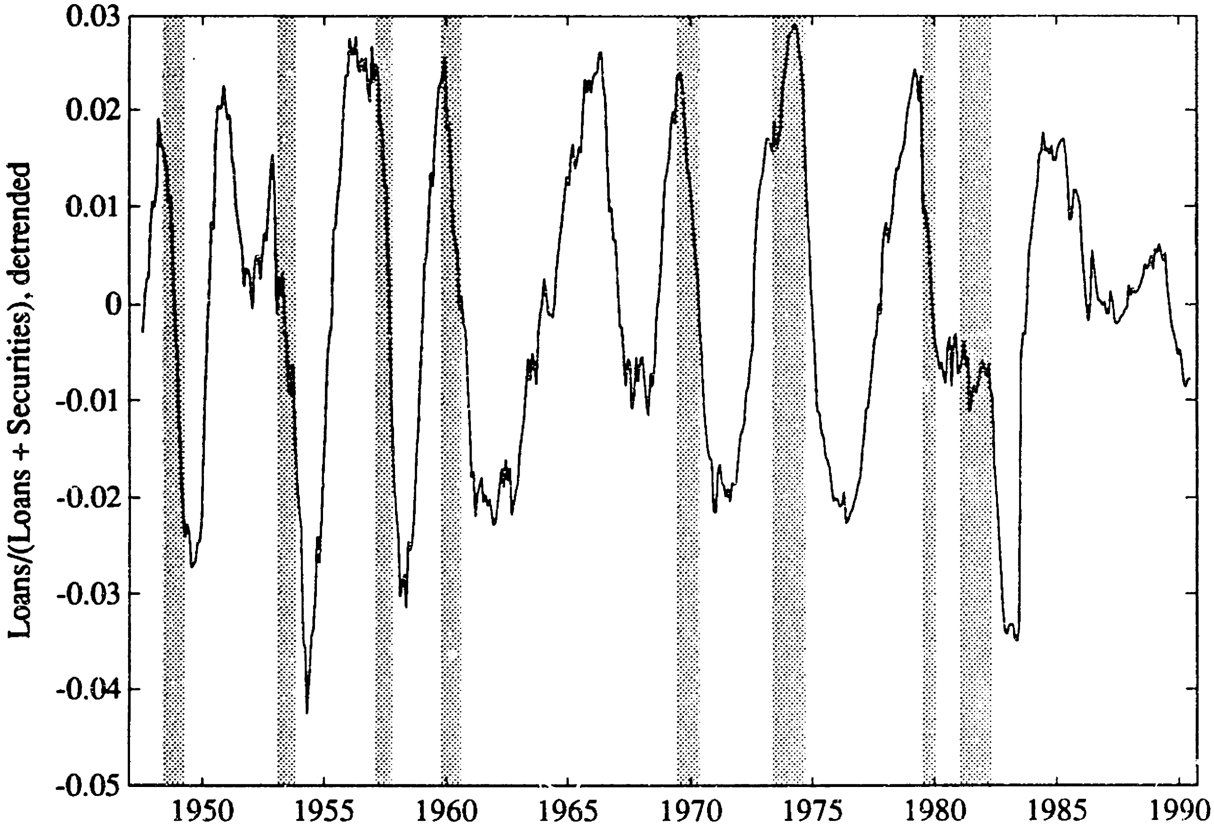



The 1957–58 and 1981–82 recessions are ambiguous because $z(t)$ was relatively stable for several months before the business cycle peak. The lags from credit to output would have to be extremely long for those two recessions to be attributed to bank lending behavior.

The two situations at a cyclical peak in figure 1-9 will be inverted at a business cycle trough. The fall in $z(t)$ during the recession could be reversed without banks granting credit to new customers if loan demand turns up or loanable funds fall, or the bank could stop the fall in $z(t)$ itself (if $z = l$) by granting credit to new customers. In the latter case, banks would again be responding nonlinearly to the incoming shocks and having an effect on the business cycle. The evidence in figure 1-10 indicates strongly that $z(t)$ keeps falling until well after the cyclical trough in all eight postwar recessions. Banks do not hit their lower bound $l$ and start adding customers until 4–8 months after the recovery has begun. This is an important difference between the model with $\gamma = 0$ and the model with $\gamma > 0$. If $\gamma = 0$, banks would not let $z(t)$ fall so far without making loans to new customers. Adding the transaction cost delays the response of bank lending to a recovery.

I would like to make two points about the preceding discussion. First, figure 1-10

Figure 1-10: $z(t)$ detrended, 1948-1990




makes clear that business cycles are different.[23] Stories such as that of Eckstein and Sinai (1986), who identify a credit crunch as a major causal factor for every postwar recession, are overstating their case. Second, the graphical evidence I have presented so far and the vector autoregression I am about to discuss depend on the relative timing of changes in bank behavior and business cycle turning points. Such timing evidence can only be suggestive, as the debate between Tobin (1970) and Friedman (1970) made clear.

[23]Blanchard and Watson (1986) make the same point.

### 1.6.5 A vector autoregression of $z(t)$ and output

The timing evidence I have just discussed is very non-statistical. Such questions of timing can also be addressed in the framework of a vector autoregression. If changes in bank lending precede business cycle fluctuations, a Granger-causality test will show that $z(t)$ leads output. If banks just react to the business cycle, output will Granger-cause $z(t)$. Table 1.1 describes the details of the vector autoregression and contains p-values for the following tests: lagged $z(t)$ has no predictive power for output, and lagged output has no predictive power for $z(t)$. The dominant relationship that showed up in figure 1-10 also shows up here. Output does precede $z(t)$ most of the time.

Table 1.1: Vector autoregression of $z(t)$ and output

| Hypothesis | P-value |
|---|---|
| Past values of $z(t)$ have no predictive power for output ($\gamma_{2i} = 0, i = 1, \ldots, s$) | 0.719 |
| Past values of output have no predictive power for $z(t)$ ($\gamma_{1i} = 0, i = 1, \ldots, q$) | 0.000 |

I ran the following vector autoregression:

$$\begin{aligned} z(t) &= \sum_{i=1}^{p} \beta_{1i}\, z(t-i) + \sum_{i=1}^{q} \gamma_{1i}\, y(t-i) + \epsilon_{1t} \\ y(t) &= \sum_{i=1}^{r} \beta_{2i}\, y(t-i) + \sum_{i=1}^{s} \gamma_{2i}\, z(t-i) + \epsilon_{2t} \end{aligned}$$

$y(t)$ is the natural logarithm of Industrial Production. I treated $y(t)$ as trend stationary and included a trend in both equations. All data was seasonally unadjusted, twelve monthly dummy variables were included in each equation, and the sample period was May 1948–August 1990. The lag lengths $p = 4$, $q = 2$, $r = 1$, $s = 1$ were chosen optimally using the Schwarz criterion. Treating $y(t)$ as difference stationary instead of trend stationary had only a small effect on the results. (Past values of $z(t)$ had more predictive power for output but the significance depended on the specific implementation of the test. When $y(t)$ was replaced by $\Delta y(t)$ and the trend was removed, the optimal lag lengths were $p = 4$, $q = 5$, $r = 1$, $s = 1$.)

## 1.7 Extensions and conclusions

### 1.7.1 If the drift of $z$ varied over time

The randomness in the model comes from the stochastic components of loan demand and assets, $dw_\lambda$ and $dw_A$; the drift terms $\mu_\lambda$ and $\mu_A$ are constant. This is not the ideal way to model business cycle fluctuations. Since $w_\lambda$ and $w_A$ are standard Brownian motions, they are martingales. Their expected change over any time horizon is zero. An important feature of business cycles is that when things are bad, it is likely that they will get better rather than worse. Business cycle randomness should not be a martingale. One way to achieve this in the model would be to vary the drift term over time in a deterministic, cyclical way. For example, let

$$\mu_z(t) = \alpha \sin(\beta t). \tag{1.4}$$

The ratio of loans to assets, $z(t)$, would follow a deterministic cycle with stochastic variation still entering through the $dw_z$ term. The period of the cycle would be $2\pi/\beta$. With a time-varying drift term the value function would depend on time. Equation (A.6) in the mathematical appendix would become

$$\rho V = f(z) + \mu_z(t) z V_z + \frac{\sigma^2}{2} z^2 V_{zz} + V_t.$$

Subscripts indicate partial derivatives. This is a second order partial differential equation and is hard to solve. A complete solution of this extension of the model is left as a project for future research. One obvious consequence is that the optimal $l$ would vary over time.

Even though I have not solved the model with time-varying drift, I can consider how the bank's optimal policy would change over the cycle if drift varied over time. From the comparative statics results in the mathematical appendix, if the growth rate of loanable funds falls or the growth rate of loan demand increases, causing $\mu_z$ to increase, the band $[l, u]$ will widen. If the drift in $z(t)$ cycles deterministically as in equation (1.4), when $\mu_z$ is high the band will be wide. As the period of high growth in $z(t)$ comes to an end, the future looks worse and worse, so the bank will be more willing to accept new customers to keep profits up during the coming bad times. The band will narrow as $l$ increases. As $\mu_z$ falls, the future will eventually begin to look better, and the band will widen. The bank will wait until $z(t)$ rises on its own rather than pay the cost of adding customers.

An optimal policy where $l$ moves over time could help account for some of the ambiguities in the previous section. For example, in figure 1-7 there are times when $z(t)$ is falling but has not quite hit bottom, yet loans made under commitment are growing so slowly relative to total loans that new customers must be getting some loans. Perhaps the optimal $l$ is falling during that time, so $z(t)$ could be falling but still be at its lower bound, implying new customers could get credit. Figure 1-7 also shows that there are times when $z(t)$ is rising from a low point and loans made under commitment are still growing more slowly than total loans. With a constant $l$, such episodes are hard to interpret, as it would appear $z(t)$ is in the interior of its band and only old customers should be getting credit. If $l$ were rising, though, $z(t)$ could still be at the bottom of its band.

### 1.7.2 Conclusion

My model of long-term customer relationships in banking is based on the informational asymmetries between banks and firms. These asymmetries make the first loan from a bank to a firm costlier than a subsequent loan, which encourages long-term relationships. Banks want to loan out all of their assets but restrain themselves because of the cost of adding new customers. New customers face tighter credit than old customers, and new customers only get credit easily when the bank's ratio of loans to assets is low. The bank's optimal behavior is to pass on exogenous shocks in a nonlinear way. If the exogenous shocks are driven by the business cycle, the bank will act differently at different phases of the cycle and the changes in bank behavior can feed back into the cycle.

A Federal Reserve Board survey of bank loan officers supported the model's conclusion that new customers face tighter credit on average. Loans made under commitment grew faster than overall lending when the ratio of loans to assets was high, as the model predicted. Looking at the response of the loans to assets ratio to monetary policy shocks highlighted the nonlinear response of bank behavior and the effect it could have on the real economy. The ratio of loans to assets shows considerable variation at the business cycle frequency, which in itself supports the model's prediction that the ratio floats within a band and is not kept at a single optimal value. An examination of the eight postwar business cycle peaks showed that in only two of the eight recessions was the nonlinear response of banks to exogenous shocks likely to have been an important factor. None of the eight recoveries could be attributed to a shift in bank lending. This exercise provided a good reminder that all business cycles should not be treated alike.

Do banks matter for business cycles? My answer is yes and no. Banks may have helped to bring on some recessions, but their influence is not as pervasive as Wojnilower (1980) or Minsky (1986) believe. In most postwar recessions and all postwar recoveries, it is unlikely that banks were a causal factor.

# Chapter 2

# Long-run Banking Relationships in General Equilibrium

## 2.1 Long-term banking relationships

### 2.1.1 Introduction

When a small business needs to borrow money to finance an investment, it turns to a bank. Banks are the best source of funds for small businesses, because producing the information needed to issue securities in capital markets would be too costly. A small business looking for credit could receive offers of credit from many banks, but it must choose to deal with one. Once a relationship is established and the bank acquires information about the small business by lending it money, the bank will be able to provide future loans at a lower cost than an outside bank. A long-term relationship between a bank and a firm takes advantage of that lower cost. The primary focus of this paper is how the contracts between the bank and the firm will be written and how these contracts affect capital accumulation.

I embed the contracting decision of the bank and the firm in a three-period overlapping generations model, with an investment technology similar to that of Bernanke and Gertler (1989). Both consumers and firms live for three periods. Because a firm lives for three periods, it makes two investment decisions: one when young and one when old. Firms must borrow to finance investment, and a long-term relationship in the model simply means the firm borrows from the same bank when old and when young.

The terms of the loan contract determine who can invest. Investment determines the next period's capital stock which determines output and wages. Consumers receive their income, consume and save, yielding a supply of funds for next period's investment. In this manner the general equilibrium of the model is established.

The model in this paper establishes that long-term banking relationships are important, but their importance varies across firms. Very productive firms bypass banks; only middle-productivity firms benefit from the lower borrowing cost a long-term relationship provides. This accords well with real-world financial arrangements. In the model, firms vary by productivity, not by size. In reality, a larger firm would be better able to bypass banks and raise funds in open securities markets because its size lets it more easily overcome the fixed costs of participating in those markets. Small firms will be more likely to deal with a bank. For this reason I motivate the model by referring to the relationship between a small business and a bank. Later in the paper I compare the loan contracts derived in the model with actual borrowing practices. The model's results are consistent with the financial arrangements observed in practice.

### 2.1.2 Related literature

Bernanke and Gertler (1989) provide a general equilibrium model where a firm's ability to get credit depends on its net worth as well as its productivity. Because theirs is a two-period overlapping-generations model, they do not consider long-term relationships. They focus on the role the net worth of firms plays in economic fluctuations. Net worth plays a role because a firm with higher net worth can supply more collateral and thereby borrow more. Their model, like mine, features costly state verification. I borrow the investment technology from Bernanke and Gertler but eliminate net worth since my focus is on long-term relationships.

Gertler (1987) considers a problem similar to mine, where a firm borrows to finance investment and project outcomes are private information, but without costly state verification. He solves for the optimal long-term contract, which is contingent on all publicly observable variables. His contract emphasizes the role of collateralizable future profits. They make a long-term contract better than short-term contracts by expanding the group of borrowers who invest. In my model firms provide no internal finance, so collateral is not an issue. My paper considers both long-term arrangements between the bank and the firm: a long-term contract, like Gertler, and a sequence of short-term contracts.

A long-term relationship is justified in this paper because it reduces monitoring costs. This is somewhat different from the approach others have taken. Sharpe (1990) assumes a bank learns about the productivity of the firms it deals with. An insider bank knows more about the productivity of the firm than an outsider bank, and the insider bank can expropriate some of the surplus from the relationship because it has a better idea of what the outcome of the firm's investment project will be. Haubrich (1989) claims that a long-term relationship can eliminate monitoring costs altogether; repeated dealings allow the bank to apply something like statistical process control techniques to the firm's output to see whether it is cheating. In an infinite horizon, by a Law of Large Numbers argument, this strategy will costlessly prevent cheating by firms. Haubrich's model, unlike mine, features moral hazard as the problem banks must deal with. One similarity is that in both models a firm builds up a reputation with its bank that cannot be transferred to another bank.

This paper explicitly considers the long-term financing relationship between a bank and a firm and puts it into a general equilibrium model where the effects of financial arrangements on capital accumulation are clear, which none of the papers just mentioned has done. Some papers have allowed for long-term relationships and others have looked at the effects of bank lending on capital accumulation, but none has considered the combination. Other models of long-term banking relationships implicitly restrict themselves to firms that deal with banks; the primary result of these models is that banks weed out poor-performing customers over time. My model considers a wider range of firms: those who deal with banks along with those too productive to need banks and those not productive enough to earn bank credit. Banks prefer to lend to middle-productivity firms, because those loans are most profitable for the bank. Banks face stiff competition for the business of high-productivity firms, who do not need banks' monitoring capabilities and can go to capital markets to finance their investments. My model has a different focus and different conclusions from other models of long-term banking relationships.

## 2.2 Setup of the model

### 2.2.1 Agents

The model economy in this paper is populated by overlapping generations of agents, who each live for three periods. Each generation has mass one and there is no population growth. All agents have a labor endowment of one in the first period of life, they care only about consumption in the third and final period of life, and they are risk neutral. The decisions of agents in this model are extremely simple: because there is no disutility of work they work when young, they save everything until their last period of life and then they consume all they have.

### 2.2.2 Production

The output and investment technology in the model economy is taken from Bernanke and Gertler (1989). Output is produced using capital and labor according to

$$y_t = f(k_t)$$

where $y_t$ and $k_t$ are output and capital per worker and $f(\cdot)$ is the production function in intensive form. Output that is not consumed is either used in an investment project (described below) or stored. Storage yields a fixed gross return $R$. Initially I assume there is always storage in equilibrium, so the equilibrium interest rate is $R$. At the end of the paper I will consider an equilibrium where this is not so.

invest $x(\omega)$ (time $t$) $\longrightarrow$ receive $\tilde{K}$ units of capital (time $t+1$)

Figure 2-1: Timing of investment projects

### 2.2.3 Firms

Along with each new generation of agents, an equally-sized cohort of firms is born. Ownership of a cohort of firms is spread evenly across the contemporaneous generation, and shares in the firms are not traded.[1] A "firm" consists of the opportunity to undertake an investment project twice, once when young and once when old. An investment project converts savings in period $t$ into a random amount of capital in period $t+1$. The amount of savings required to fund a project varies across firms, but the distribution of the random amount of capital produced is the same for all firms and is not dependent on firm effort. This implies there will be neither moral hazard nor adverse selection in the model.

An investment project requires input $x(\omega)$ and yields a random amount of capital $\tilde{K}$. See Figure 2-1. $\tilde{K}$ is distributed continuously on $[\underline{K}, \overline{K}]$ with mean $K$ and density function $g_K$. The distribution of $K$ is common knowledge and exhibits increasing hazard.[2] Firms are indexed by the parameter $\omega$, uniform on $[0,1]$, which is public information. A firm with productivity $\omega$ requires $x(\omega)$ units of savings to undertake its project. I assume $x' > 0$, so a firm with higher $\omega$ requires more savings to fund its project. Since all projects have the same probability distribution $g_K$ (although not the same outcome!), a firm with higher $\omega$ is less efficient, *ex ante*. A firm with high $\omega$ will be referred to as a low productivity firm (and vice versa) throughout the paper. All investment projects are independent draws from the distribution $G_K$, so there is no aggregate uncertainty about the total amount of capital produced by any (large) number of investment projects. (It is $K$, the mean of $\tilde{K}$, times the number

[1]Ownership in this model economy simply means the right to a fraction of the firm's profits, if any. It does not confer any right of control or any liability for covering the firm's losses should it default.

[2]Increasing hazard is a common assumption in the financial contracting literature. See, for example, Morgan (1992).

of projects.) A firm will operate its project if it can obtain financing and would earn positive expected profits by doing so. Firms are risk-neutral. Any profit realized from the firm's investment is distributed among the agents in the contemporaneous generation.

### 2.2.4 Information

The outcome of a project is known only to the firm undertaking the project. Other agents in the economy are able to verify the outcome of the project, but only by incurring a cost. The information structure in the economy is one of "costly state verification" (Townsend 1979).

If an agent monitors a firm's project outcome, the outcome is known only to that agent.[3] If many agents want to know the result of a single firm's investment project, they must each pay the verification cost. To avoid duplication of monitoring, zero-profit intermediaries ("banks") will emerge to do all the lending and monitoring, as in Diamond (1984). These intermediaries can guarantee a fixed return to depositors while investing in risky projects because they diversify their lending across a large number of firms. Unlike agents and firms, banks will be long-lived. In the period after the loan is made, the capital produced by the investment project is divided between the bank and the firm according to the terms of their loan contract.

### 2.2.5 Monitoring costs

To provide a motivation for long-term interaction, I assume that monitoring the outcome of a firm's investment project is more costly for a first-time lender than for a repeat lender. To minimize monitoring costs, a firm investing for the second time will borrow from the same bank it borrowed from the first time. I assume that monitoring costs $\gamma_1 + \gamma_2$ the first time a bank lends to a firm, but only $\gamma_1$ the second time, because the bank learns something about the firm by dealing with it for a period. Even if no monitoring occurs in the first period of the relationship, the firm still benefits from

[3]This is in contrast to Bernanke and Gertler, where the results of monitoring are made public.

lower monitoring costs in the second period. because the bank has learned something about the firm in dealing with it the first time, whether or not monitoring took place.[4]

## 2.3 Contracts between banks and firms

A firm goes to a bank to get the credit it needs to invest. To each firm that asks for credit, a bank will offer a loan contract, consisting of a loan amount and a repayment schedule. Repayment can take place in either or both periods of the firm's life and can depend on the outcomes of the firm's projects when young and old. The terms of the contract between bank and firm will differ depending on whether they sign a single long-term contract when the firm is young or a pair of short-term contracts, one when young and one when old. A firm will borrow from only one bank; to borrow from more than one would be more expensive because it would require duplication of monitoring. In the remainder of this section, I discuss the potential contracts and their implications for aggregate investment. I compare the results under the various contracts.

### 2.3.1 Perfect information

As a benchmark, consider the firm's investment decision under perfect information. This case would apply if the outcome of the firm's investment project were public information, if state verification were costless or if the firm had internal funds to invest. Under perfect information, the firms whose investment projects have positive expected value, net of opportunity cost, would operate. Other firms would not operate and relationships between banks and firms would not be an issue. The expected amount of capital produced by an investment project will be $K$ (the expected value of $\tilde{K}$). If the expected price of capital next period is $q$, the expected output of the project is $qK$. The opportunity cost of undertaking the investment is just the interest

[4]One way to justify this structure of monitoring costs, based on Fama (1985), would be to assume firms maintain deposit balances during the first period of a relationship. Holding a firm's deposits gives a bank valuable information that reduces monitoring costs.

lost on the savings invested in the project: $Rx(\omega)$. The social surplus is $qK - Rx(\omega)$. I assume the most productive firm's investment is always socially profitable:

$$qK - Rx(0) > 0,$$

and the least productive firm's investment is never socially profitable:

$$qK - Rx(1) < 0,$$

for all possible $q$, to rule out a corner solution. All projects with $qK - Rx(\omega) > 0$ will be undertaken. If $\tilde{\omega}$ is defined by $qK - Rx(\tilde{\omega}) = 0$, firms with $\omega \leq \tilde{\omega}$ have socially profitable investments and firms with $\omega > \tilde{\omega}$ do not.

### 2.3.2 An assortment of contracts

I will consider three types of interaction between the bank and the firm in what follows. All three consist of two one-period debt contracts. They differ in the link between the terms of the loan when the firm is young and the terms of the loan when the firm is old. The first case is the unrealistic one of no link, and is described solely as a benchmark. The second is a long-term *contract*: terms of both loans are fixed before the first-period loan is made. The third is a long-term *relationship*: the firm deals with the same bank when young and when old, but the terms of its second-period loan are not determined until the second loan is made. Before proceeding I summarize the one-period debt contract, the building block upon which all three contracts are based.[5]

Under a debt contract, the firm repays a constant amount $F$ to the bank, unless its project has a bad outcome and yields less than $F$, in which case the bank pays to monitor the firm's project outcome and takes everything the firm produced.[6] A

[5]This is the contract first derived by Townsend (1979) which has since become a standard in the financial intermediation literature.

[6]The contractual interest rate $i$ is related to the contractual repayment $F$ by

$$(1+i)\,x(\omega) = F.$$

firm is only monitored if its investment turns out badly. For the bank to earn zero expected profit net of opportunity cost, the payment to the bank ($F$) must satisfy

$$q\left[\int_{\underline{K}}^{F}(K-\gamma)g_K(K)dK+\int_{F}^{\overline{K}}F\,g_K(K)dK\right]=Rx(\omega) \tag{2.1}$$

where $q$ is the price of capital next period, $R$ is the interest rate, $\gamma$ is the monitoring cost and $x(\omega)$ is the amount a firm must borrow to operate its project. The left-hand side of (2.1) is the expected repayment to the bank and the right-hand side is the opportunity cost of the funds invested in the project. Figure 2-2 shows the bank's expected repayment (the left-hand side of (2.1)) as a function of $F$, the contractual repayment. As points A and B in the figure show, lower productivity firms (with higher $\omega$) must pay a higher interest rate to receive a loan. A debt contract has a maximum contractual repayment of $\overline{F}$ above which the bank's expected repayment falls. No contract will call for a repayment greater than $\overline{F}$. $\overline{F}$ is defined by the first-order condition that results from maximizing the left-hand side of (2.1) with respect to $F$:

$$-\gamma g_K(\overline{F})+(1-G_K(\overline{F}))=0$$

or

$$\frac{g_K(\overline{F})}{1-G_K(\overline{F})}=\frac{1}{\gamma} \tag{2.2}$$

Because $\overline{F}$ is a function of $\gamma$, it will be written as $\overline{F}(\gamma)$ from now on. Firms with high enough $\omega$ will not be granted a loan. For future reference, notice that the expected repayment and the cutoff $\overline{F}$ both are lower when the monitoring cost $\gamma$ is higher.

### 2.3.3 Myopic long-term relationship

As a benchmark against which to measure the contracts described in the next two sections of the paper, this section describes a myopic long-term relationship. The adjective "myopic" applies because the parties make a loan arrangement when the firm is young without looking ahead to the next period and realizing the firm will

In the rest of the paper I use "repayment" and "interest rate" interchangeably.

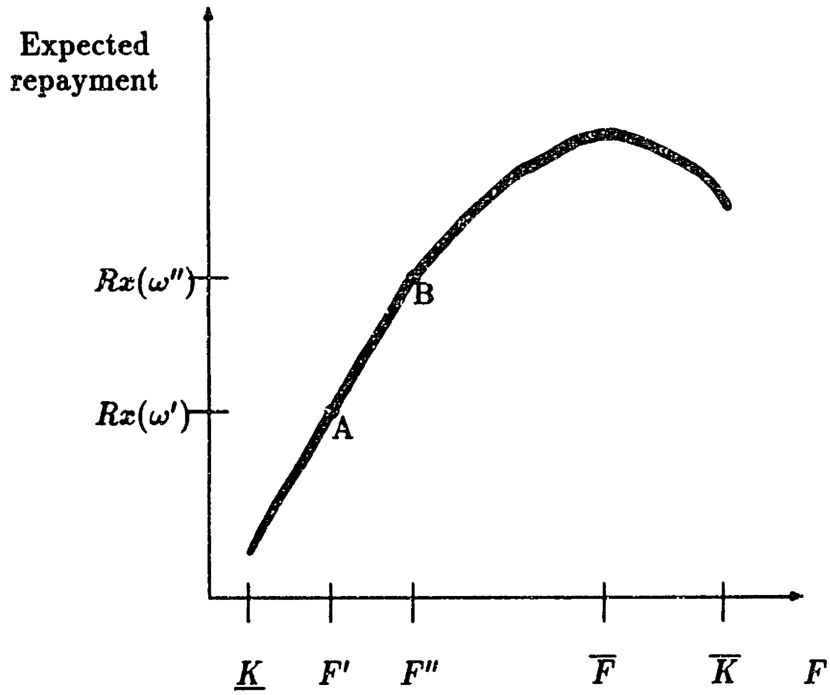



Figure 2-2: Bank's expected repayment as a function of contractual repayment

also want credit when old. In this scenario, in both the first and second periods of its life a firm obtains a bank loan at the interest rate that earns the bank zero profits in that period.

The two interest rates $F^1$ and $F^2$ are determined by

$$q_{t+1}\left[\int_{\underline{K}}^{F^1}(K-(\gamma_1+\gamma_2))g_K(K)dK+\int_{F^1}^{\overline{K}}F^1\,g_K(K)dK\right]=Rx(\omega) \qquad (2.3)$$

and

$$q_{t+2}\left[\int_{\underline{K}}^{F^2}(K-\gamma_1)g_K(K)dK+\int_{F^2}^{\overline{K}}F^2\,g_K(K)dK\right]=Rx(\omega), \qquad (2.4)$$

which is just the one-period zero profit condition (2.1) rewritten for the two periods in which a firm born at time $t$ is alive. Neither party takes account of the long-run nature of their interaction.

The contract will be feasible for firms with $\omega$ low enough so that in both periods an interest rate exists that earns the bank zero profit ($F^1 \leq \overline{F}(\gamma_1+\gamma_2)$ and $F^2 \leq \overline{F}(\gamma_1)$).

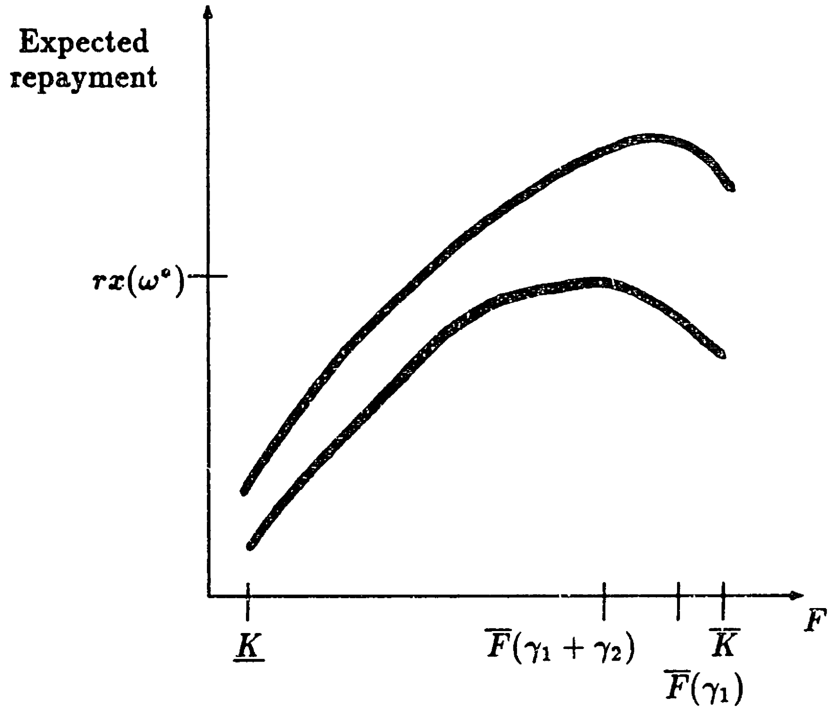



Figure 2-3: Interest rate determination in the myopic long-term relationship

Since monitoring costs are higher in the first period, the first period's loan will break down first. All firms with $\omega < \omega^*$ will get credit, where $\omega^*$ is defined by

$$q_{t+1}\left[\int_{\underline{K}}^{\overline{F}(\gamma_1+\gamma_2)}(K-(\gamma_1+\gamma_2))g_K(K)dK+\int_{\overline{F}(\gamma_1+\gamma_2)}^{\overline{K}}\overline{F}(\gamma_1+\gamma_2)\,g_K(K)dK\right]=Rx(\omega^*)$$

See Figure 2-3. The cutoff $\omega^*$ marks the firms that would receive credit and invest if banks ignored the fact that a firm lives for two periods.

### 2.3.4 Long-term contract

#### What is a long-term contract?

The second of the three contracts to be analyzed is a long-term contract, which could also be described as the optimal contract. A sequence of short-term contracts where the terms are agreed upon in advance is one type of long-term contract, but it will not necessarily be the optimal long-term contract. The optimal long-term contract could take a more complex form. Townsend (1979) makes an effort to characterize the optimal contract in an environment with more than one random variable that is subject to costly state verification. By restricting exchanges to be bilateral, along with other technical assumptions, he proves that each individual (period's) contract takes the simple form described by (2.1) above. However, the bilateral exchange assumption is clearly violated here, because the two agents whose outcomes are subject to costly state verification are the young firm and the old firm, and the young firm's behavior could (and perhaps should) affect the old firm's loan contract. I conjecture that the optimal long-term contract would be more complex than two short-term contracts with precommitted loan amounts and repayment terms. Still, that is the case considered below.

### Terms of the contract

The long-term contract analyzed here is made up of two one-period debt contracts. The interest rates in both periods are agreed upon in the first period. Right away we can single out the most productive firms, those who can repay what they borrow in every state of nature. Such a firm will never be monitored, so the lower second-period monitoring cost is immaterial. Since lending is channeled through banks to avoid duplication of monitoring, if monitoring is certain not to occur firms can borrow directly from a group of savers, bypassing banks altogether. The interest rate paid by the firm satisfies $qF = Rx(\omega)$, which is just the lender's zero-profit condition. These firms will have $\omega \leq \underline{\omega}$, where $\underline{\omega}$ is defined by $q\underline{K} = Rx(\underline{\omega})$. In what follows the loan contracts of the remaining firms, with $\omega > \underline{\omega}$, are considered.

If, when they initiate a two-period relationship, the bank and firm agree to interest rates and loan amounts for both periods of the relationship, the following contract would emerge. The bank earns no expected profits over the life of the relationship, but it could earn positive profit in one period, balanced by a loss in the other period. The optimal contract would choose $F^1$ and $F^2$ to minimize monitoring costs,

$$q_{t+1}(\gamma_1 + \gamma_2)G_K(F^1) + q_{t+2}\,\gamma_1\,G_K(F^2) \qquad (2.5)$$

while earning zero profits for the bank:

$$q_{t+1}\left[\int_{\underline{K}}^{F^1}(K-(\gamma_1+\gamma_2))g_K(K)dK+\int_{F^1}^{\overline{K}}F^1 g_K(K)dK\right] \tag{2.6}$$

$$+\,q_{t+2}\left[\int_{\underline{K}}^{F^2}(K-\gamma_1)g_K(K)dK+\int_{F^2}^{\overline{K}}F^2 g_K(K)dK\right]=2\,Rx(\omega).$$

To minimize (2.5) subject to (2.6), form a Lagrangian and take first-order conditions.[7] The first-order conditions are

$$(\gamma_1+\gamma_2)g_K(F^1)(1-\lambda)+\lambda(1-G_K(F^1))=0 \tag{2.7}$$

$$\gamma_1\, g_K(F^2)(1-\lambda)+\lambda(1-G_K(F^2))=0 \tag{2.8}$$

where $\lambda$ is the Lagrange multiplier on (2.6). Combining (2.7) and (2.8) gives

$$\frac{h_K(F^1)}{h_K(F^2)}=\frac{\gamma_1}{\gamma_1+\gamma_2} \tag{2.9}$$

where $h_K$ is the hazard function associated with the density $g_K$.[8] Since I have already assumed $h_K$ is increasing, (2.9) proves that $F^1 < F^2$. The bank charges the firm less in the first period of the relationship than in the second because monitoring costs are higher in the first period, and the bank and firm commit to work together to reduce monitoring costs. This motivation for a low introductory rate is very different from what we will see in the next section. Combining (2.9) with the bank's zero profit condition yields the solutions for $F^1$ and $F^2$. See Figure 2-4, where the first-order condition (2.9) is labeled "FOC" and the bank's outside return constraint (2.6) is labelled "ORC." The intersection, point B, reveals the optimal choice of $F^1$ and $F^2$.

The interest rates $F^1$ and $F^2$ depend on $\omega$, as we see from (2.6), the bank's zero profit condition. As $\omega$ rises, the expected repayment to the bank increases to cover the opportunity cost of more borrowed funds. In Figure 2-4, ORC shifts out as $\omega$

[7] A complete statement of the minimization would include two inequality constraints on $F^1$ and $F^2$, that they both be greater than or equal to $\underline{K}$. These are omitted in the text for expositional simplicity.

[8] $h_K(K) \equiv g_K(K)/(1-G_K(K))$.

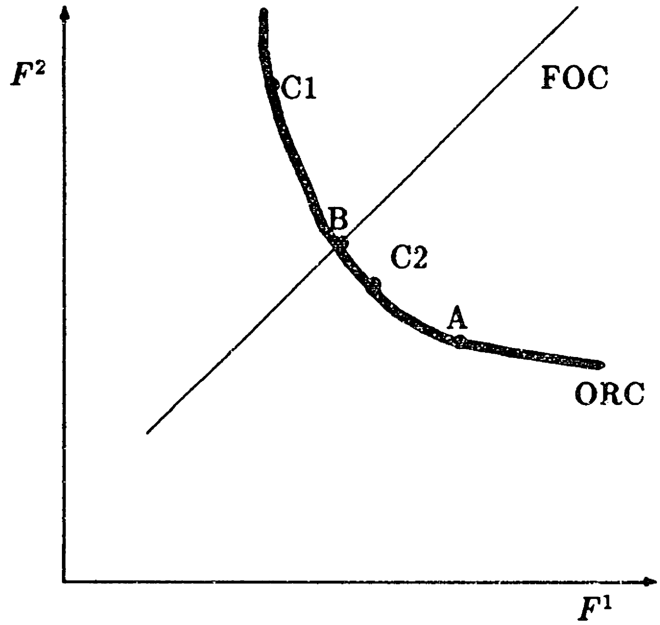



Figure 2-4: Optimal interest rates in the three contracts

increases, and $F^1$ and $F^2$ will both increase. At some level of $\omega$, it becomes impossible for the bank to recoup its investment; once $F^1 = \overline{F}(\gamma_1 + \gamma_2)$ and $F^2 = \overline{F}(\gamma_1)$ the expected repayment to the bank is at its maximum.[9] Above this level of $\omega$, the contract breaks down. Call this level $\hat{\omega}$.

Firms with $\omega \leq \hat{\omega}$ will receive an offer of credit from a bank. Which firms accept the offer and choose to operate? Any firm with positive expected profit will do so. Let

$$\mathcal{F}(F^1, F^2) = q_{t+1}\left[\int_{F^1}^{\overline{K}}(K - F^1)g_K(K)dK\right] + q_{t+2}\left[\int_{F^2}^{\overline{K}}(K - F^2)g_K(K)dK\right]$$

be the expected profit of the firm. Firm profits are always positive for any firm that operates its investment project, because the firm risks none of its own funds and the repayments $F^1$ and $F^2$ will always be strictly less than $\overline{K}$. Note that firm profits fall as $\omega$ rises because $F^1$ and $F^2$ both increase. Still, $\mathcal{F} > 0$, so every firm with $\omega \leq \hat{\omega}$ will ask for and receive credit from a bank in both periods of its life.

When operating under a long-term contract, the bank and firm agree up front what

[9]To see that $F^1 = \overline{F}(\gamma_1 + \gamma_2)$ and $F^2 = \overline{F}(\gamma_1)$ simultaneously, combine (2.9) with (2.2).

the present and future interest rates should be. However, if no binding commitment mechanism exists, the bank could not commit to a second period interest rate. Once the second period of the relationship begins, the bank has some market power that it could exploit, and it will no longer be optimal for the bank to go along with the agreement it made.[10] I now turn to the contract that would emerge in that case.

### 2.3.5 Long-term relationship

A different contract will be written if the second period loan is not contractible in the first period of the relationship. The bank that supplied credit to the firm in the first period (the "insider" bank) will be able to supply it with credit in the second period at a lower cost, because the information it gathered by dealing with the firm in the first period allows it to monitor the firm more cheaply, should monitoring be necessary. In the second period of the relationship, the bank will exploit the market power its informational advantage over the other banks gives it. In the first period, the bank and the young firm will negotiate a loan, both aware the bank will have an advantage over other banks once the relationship has been established. The bank is unable to commit not to exercise its market power, as it had to be able to do to implement the long-term contract.

To find the details of the sequence of short-term contracts, I work backwards from the second period, as is usual in such problems. First, I find the profit-maximizing interest rate charged by the insider bank in the second period. The bank will earn some profit in the second period of the relationship. Going back to the first period of the relationship, the bank will offer an interest rate that loses it enough money that the total profit from the relationship, first and second periods together, is zero, as it must be since banking is competitive.

[10]I ignore reputation effects, which could be important here. For a discussion of reputation, see footnote 16.

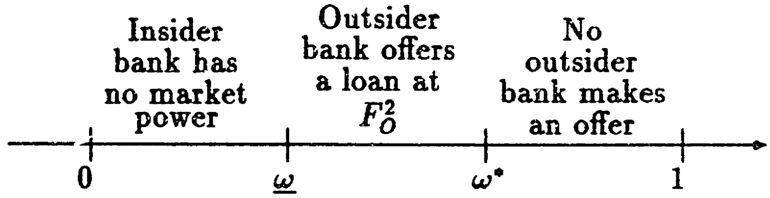



Figure 2-5: Second period offers

### Terms of the second period loan

Recall that firms with $\omega \leq \underline{\omega}$ are never monitored because they borrow so little they are able to repay in every state of nature. Monitoring costs are irrelevant for these firms, and the insider bank has no advantage over other banks when competing for these firms' second-period business.

For firms with $\omega > \underline{\omega}$, expected monitoring costs are positive so the cost advantage of the insider bank is relevant. In the second period of the relationship, an outsider bank offers to lend $x(\omega)$ to the firm at an interest rate that yields it zero profit. Call this interest rate $F_O^2$; it satisfies

$$q_{t+2}\left[\int_{\underline{K}}^{F_O^2}(K-(\gamma_1+\gamma_2))g_K(K)dK+\int_{F_O^2}^{\overline{K}} F_O^2\, g_K(K)dK\right]=Rx(\omega). \qquad (2.10)$$

For $\omega$ above some cutoff $\omega^*$, this equation has no solution (because at $\omega^*$, $F_O^2 = \overline{F}(\gamma_1+\gamma_2)$). Firms less productive than $\omega^*$ are not able to get a second-period loan from an outsider bank because the outsider bank can not charge an interest rate high enough to cover the cost of making the loan.

Firms with $\omega > \underline{\omega}$ can be divided into two groups: those who receive an offer from an outsider bank ($\omega \in (\underline{\omega}, \omega^*]$) and those who do not ($\omega > \omega^*$). See Figure 2-5. The insider bank treats the two groups differently. For firms with an outside offer, the insider bank can retain the customer by charging an interest rate less than or equal to $F_O^2$. Because it maximizes profits, it always charges an interest rate equal to $F_O^2$. The insider bank limit prices against the outsider bank's offer. Let $\mathcal{F}_2$ denote the firm's profits in the second period. Those profits are

$$\begin{aligned}\mathcal{F}_2(F_O^2) &= q_{t+2}\int_{F_O^2}^{\overline{K}}(K - F_O^2)g_K(K)dK \\ &= q_{t+2}K - Rx(\omega) - q_{t+2}(\gamma_1 + \gamma_2)G_K(F_O^2).\end{aligned}$$

The bank's second-period profits ($\mathcal{B}_2$) are

$$\mathcal{B}_2(F_O^2) = q_{t+2}\,\gamma_2\,G_K(F_O^2).$$

Looking back to (2.10), note that $dF_O^2/d\omega > 0$, which implies $d\mathcal{B}_2/d\omega > 0$ and $d\mathcal{F}_2/d\omega < 0$.

For firms without an outside offer, the insider bank acts as a monopolist and charges its revenue-maximizing interest rate $\overline{F}(\gamma_1)$. The bank is the residual claimant in this region. The firm is left with

$$\mathcal{F}_2(\overline{F}(\gamma_1)) = q_{t+2}\int_{\overline{F}(\gamma_1)}^{\overline{K}}(K - \overline{F}(\gamma_1))g_K(K)dK$$

while the bank earns

$$\mathcal{B}_2(\overline{F}(\gamma_1)) = q_{t+2}K - Rx(\omega) - q_{t+2}\,\gamma_1\,G_K(\overline{F}(\gamma_1)) - \mathcal{F}_2(\overline{F}(\gamma_1)).$$

In this region, $d\mathcal{B}_2/d\omega < 0$ and $d\mathcal{F}_2/d\omega = 0$. For $\omega$ large enough, the bank's second-period profit becomes negative and the bank does not make a second-period loan. Call this cutoff $\overline{\omega}$.

To summarize, in the second period of a relationship, four things could happen (see Figure 2-6):

$\omega \leq \underline{\omega}$ The firm borrows so little that it repays the loan in every state of nature. Monitoring never occurs so the insider bank has no advantage over other banks. The insider bank can make no second-period profits: $\mathcal{B}_2 = 0$. The interest rate $F^2$ satisfies a zero-profit condition for the bank: $qF^2 = Rx(\omega)$.

$\underline{\omega} < \omega \leq \omega^*$ The firm borrows enough that expected monitoring costs are positive,

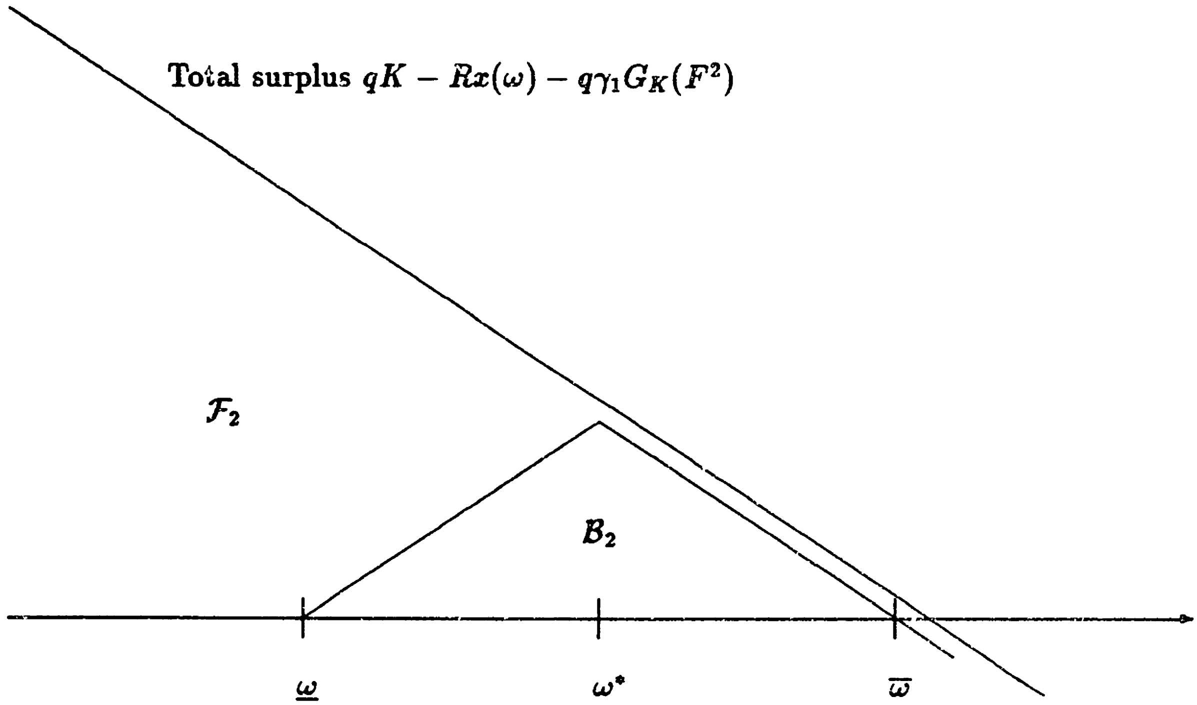



Figure 2-6: Second period bank profits

but not so much that no outsider bank makes it an offer. An outsider bank offers the firm a loan that leaves the outsider bank with zero profits and the firm with positive profits. The insider bank limit prices against the outsider bank's offer, giving the firm the same profit and keeping the savings in reduced monitoring costs for itself. $\mathcal{B}_2 > 0$. $F^2 = F_O^2$.

$\omega^* < \omega \leq \overline{\omega}$ The firm borrows so much that expected monitoring costs are high enough to make a potential relationship with an outsider bank unprofitable for the bank. The insider bank is not constrained by outside competition and extracts all the surplus it can from the relationship. The firm is left with some profit since $\overline{F}(\gamma_1) < \overline{K}$. $\mathcal{B}_2 > 0$. $F^2 = \overline{F}(\gamma_1)$.

$\omega > \overline{\omega}$ The firm needs to borrow so much that even a loan from an insider bank would be unprofitable. The firm cannot get credit and does not operate in the second period. $\mathcal{B}_2 = 0$.

I have characterized the second-period loan contract between a firm of any productivity and its insider bank. Now I turn to the first period of the relationship.

**Terms of the first-period loan**

When a bank and a young firm meet to negotiate a loan, both know what will happen when the firm is old and is a "captive customer" of the bank. In other words, $B_2$ is known. The bank offers the young firm an interest rate $F^1$ that ensures the bank earns zero profit over the course of the relationship ($B_1 + B_2 = 0$).[11] If $B_2 > 0$, this implies $B_1 < 0$, so the bank makes a loss on its first period loan. $F^1$ is determined by

$$q_{t+1}\left[\int_{\underline{K}}^{F^1}(K-(\gamma_1+\gamma_2))g_K(K)dK + \int_{F^1}^{\overline{K}} F^1 g_K(K)dK\right] = Rx(\omega) - B_2 \qquad (2.11)$$

As $\omega$ increases, $x(\omega)$ increases and $B_2$ changes, first increasing because outside competition becomes weaker and then decreasing because the bank is taking almost all of the surplus from the relationship and the surplus is falling (see Figure 2-6). At some level of $\omega$ between $\omega^*$ and $\overline{\omega}$, the bank becomes unable to earn zero profit on its funds; (2.11) will have no solution because $F^1 > \overline{F}(\gamma_1 + \gamma_2)$. That level of $\omega$ will be $\hat{\omega}$. A firm with $\omega \leq \hat{\omega}$ will invest in both periods of its life. This cutoff is identical to that in the long-term contract, so the same group of firms receives credit in both financial arrangements. The marginal firm ($\omega = \hat{\omega}$) must be charged $F^1 = \overline{F}(\gamma_1 + \gamma_2)$ and $F^2 = \overline{F}(\gamma_1)$ for the bank's outside return constraint to be satisfied, regardless of whether a long-term relationship of a long-term contract is in force. Comparing these two arrangements with the perfect information case, fewer firms get credit ($\hat{\omega} < \check{\omega}$) because the costs of monitoring make some firms unprofitable.

Figure 2-6 reveals that middle-productivity firms gain the most from a long-term relationship. Firms with productivity near $\omega^*$ yield the most second-period profit to the bank, and all second-period bank profits are transferred to the firm via a low first-period interest rate. Middle-productivity firms need banks because some information about their investments must be generated in order to make a loan feasible, and banks generate that information at lowest cost. Very productive firms ($\omega < \underline{\omega}$) gain nothing from a long-term relationship because they can finance their investments without agency costs. The publicly available information on these firms ($\omega$ and $\underline{K}$)

[11] For simplicity, I assume neither the firm nor the bank discounts the future.

is sufficient to guarantee lenders a competitive return with no chance of monitoring, eliminating the need for bank intermediation entirely. Very unproductive firms ($\omega > \overline{\omega}$) gain nothing because even though their investments may have positive social value ignoring monitoring costs (if $\omega \in (\overline{\omega}, \check{\omega})$), the monitoring is essential to produce enough information about the investment outcome to elicit the necessary credit. Even the lower monitoring cost of a two-period relationship cannot make these firms' investments profitable. Internal finance, through firm equity, would help all firms and would expand the range of firms that can operate by reducing the need for external funds which are accompanied by a risk of monitoring.[12]

### 2.3.6 Interpretation : Compare and contrast the three

I have discussed three possible arrangements between a bank and a firm in this model economy: a myopic long-term relationship, a long-term contract, and a long-term relationship. Each would provide a firm with a different pair of interest rates and a different level of profit. The same group of firms (those with $\omega \leq \hat{\omega}$) will get credit under the last two arrangements; fewer firms would get credit if the parties were myopic.[13]

We can first compare the myopic long-term relationship with the long-term contract. In Figure 2-4, point A represents the interest rate pair of the former and point B the latter. Both lie on the ORC curve, since the bank makes zero profit in both cases. The long-term contract conserves on monitoring costs by concentrating more of the monitoring in the second period, where it is less costly. The myopic long-term relationship does the opposite because the bank breaks even in each individual period. In a one-period debt contract, higher monitoring costs require a higher interest rate. Monitoring costs more in the first period, so the first-period interest rate exceeds the second-period interest rate under a myopic long-term relationship. More monitoring occurs in the first period, where it is relatively expensive. Not only is monitoring

[12]This is the main point of Bernanke and Gertler (1989).

[13]If firms could finance their investments with inside funds, as they do in Bernanke and Gertler (1989), the set of firms that get credit under the long-term relationship and the long-term contract could be different.

more expensive, making investment less productive, but fewer firms are able to invest because the cutoff for getting credit is lower ($\tilde{\omega} < \hat{\omega}$) under a myopic long-term relationship. More capital is produced and firm profits are larger in an economy with long-term contracts.

The more realistic comparison is between the long-term contract and the long-term relationship. In Figure 2-4, these are points B and C respectively. (Point C can lie either above or below the FOC line, as represented by points C1 and C2.) Point C, representing the long-term relationship, must lie above the 45 degree line, since the bank opportunistically charges a high interest rate in the second period. However, point C can lie on either side of the FOC line. For low levels of $\omega$, the bank will not have much market power in the second period and $F^1$ and $F^2$ will be nearly equal and close to $\underline{K}$. $F^1$ will always be lower than $F^2$ (if $\omega > \underline{\omega}$) so the bank can make a first-period loss to counter its second-period profit. The amount of capital wasted by monitoring will be greater at C than at B, simply because B minimizes monitoring costs. Again, firm profits will be higher and more capital will be produced under a long-term contract than under a long-term relationship despite the fact that the same group of firms operate in both cases. Financial intermediation is more efficient if banks and firms write long-term contracts; long-term relationships result in more resources being devoted to monitoring, because the interests of the bank diverge from the interests of the firm.

Whether the long-term contract or the long-term relationship emerges depends on whether the second-period interest rate is contractible in the first period the bank and the firm deal with each other. In the very simple model economy described so far, nothing prevents the bank and the firm from contracting on the second period interest rate. There is no aggregate uncertainty in the model, so the parties can compute their expected return from any potential future interest rate. If the model economy were made more complex, the second-period interest rate could become not contractible. In the contracts literature, one reason for a variable to be considered not contractible is if economic conditions are expected to change in such an unpredictable way that agents cannot form an expectation with which to compute their payoffs. Or, it could be too costly to compute such an expectation. In the present case, the uncertainty of the real world has been vastly simplified. If more real-world complexity were allowed into the model, the second-period interest rate might become not contractible. The model in this paper implies the long-term contract will be chosen if uncertainty is well-defined and well-understood; a more complex model might lead to the opposite deduction.

A second way to establish which is the more relevant contract is to look at real-world financial arrangements to see whether they more closely resemble the long-term contract or the long-term relationship. The long-term contract as defined here shares some features with the loan commitments that have become an increasingly important arrangement for U.S. commercial banks and their loan customers.[14] Both specify an interest rate and loan amount in advance. For the analogy to the model to hold, however, the loan commitment would have to last for the duration of the relationship, which may not accord well with the real world. If a bank and a firm renegotiate their loan commitment, the long-term relationship becomes relevant. If the bank cannot precommit not to use its market power upon renegotiation, the long-term contract is infeasible. Comparing the results of the model with the stylized facts of the U.S. banking system leads to the conclusion that such precommitment does not occur, so by induction the long-term relationship is the contract of choice. Since monitoring costs are higher under the long-term relationship, the fact that real-world financial arrangements resemble the long-term relationship implies the costs that are not modeled, such as the cost of actually writing the contract, are high enough under the long-term contract to make it the more expensive of the two when all costs are considered.

The long-term contract and the long-term relationship feature a low first-period interest rate, but for different reasons. The bank charges a low interest rate when initiating a long-term contract because monitoring is more expensive in the first period and both parties strive to minimize monitoring costs. When a bank enters

[14]According to the March 1992 *Federal Reserve Bulletin*, 74 percent of short-term commercial bank lending is done under a loan commitment. For a model of loan commitments, see Morgan (1992).

into a long-term relationship, it charges a low interest rate to capture the firm as a customer; it knows it will obtain some market power and can earn some rents in the future.[15] Both the long-term contract and the long-term relationship contrast with the myopic long-term relationship, where the interest rate is high in the first period and low subsequently.

The choice between a long-term contract and a long-term relationship comes about because of a conflict of interest between the bank and the firm. If the bank acts solely according to self-interest, it will charge as high an interest rate as it can in the second period of the relationship. If the bank and the firm work together to maximize joint profit, the long-term contract would be feasible.[16] This distinction suggests an international comparison of banking practices. In Anglo-Saxon countries (the United States and United Kingdom), ownership of banks and firms is separated by law. In Europe and Japan, banks have close ties to the firms they lend to, often including an equity stake.[17] The latter arrangement, called universal banking, ensures the bank will not act out of pure self-interest, because it reaps some of the benefits of its customers' profits. The financial intermediation sector in universal banking countries may be closer to long-term contracts, while intermediation in Anglo-Saxon countries may be closer to long-term relationships.[18] Of course, as discussed above, this has a direct implication for the efficiency of financial intermediation in the two groups of countries. Universal banking would be more efficient, if the conclusions of the model could be applied this way, because there would be less dead-weight loss from the monitoring of borrowers by lenders.

[15]This is the motivation for a low introductory price typically found in models of repeated interaction with relationship-specific investment, such as Farrell and Shapiro (1989) and Sharpe (1990).

[16]The bank's eagerness to gouge the firm in the second period obviously ignores considerations of reputation. Sharpe (1990) discusses the effects of reputation. He finds that reputation can mitigate the problems of bank opportunism in the second period of a relationship. In his model with reputation, the outcome under a long-term relationship can duplicate the outcome under a long-term contract if certain technical conditions hold.

[17]See Bisignano (1990) for a comparison of the two systems.

[18]If the bank and the firm were one and the same, the need for costly state verification would be eliminated as the bank could provide costless "internal" funds to finance the firm's investment.

## 2.4 General equilibrium

Given an agreement between banks and firms on the terms of financing, investment and capital accumulation can be determined. The next step is to calculate the equilibrium over time, as successive cohorts of firms and agents are born, live, and die. The general equilibrium of the model is determined exactly as in Bernanke and Gertler (1989). At the beginning of period $t$, the capital stock $k_t$ is predetermined (by last period's investment decisions). The capital stock determines output and the wage. Firms that invested in $t-1$ sell whatever capital their investments netted them and redistribute profits to their owners. Saving is equal to the wages of the young plus the saving of the middle aged. I assume for now that saving is always adequate to fund all projects with positive social value, so rationing never occurs. Investment remains to be determined.

Under either long-term contracts or long-term relationships, every firm with $\omega < \hat{\omega}$ will invest; the cutoff level $\hat{\omega}$ depends positively on $q_{t+1}$ and $q_{t+2}$, so $\hat{\omega}$ can vary with $t$ (and will be written as $\hat{\omega}_t$ where necessary to avoid confusion). Firms born in $t$ with $\omega \leq \hat{\omega}_t$ will invest in period $t$; so will firms born in $t-1$ with $\omega \leq \hat{\omega}_{t-1}$. Supply of next period's capital stock $k_{t+1}$ is equal to the sum of capital produced by old firms and capital produced by young firms.

The investments of young firms in period $t$ yield

$$\int_0^{\hat{\omega}_t} (K - (\gamma_1 + \gamma_2) G_K(F^1(\omega)))\, d\omega$$

units of capital in $t+1$; those of old firms yield

$$\int_0^{\hat{\omega}_{t-1}} (K - \gamma_1\, G_K(F^2(\omega)))\, d\omega$$

and the interest rates $F^1$ and $F^2$ have been written explicitly as functions of $\omega$ as a reminder that less productive firms pay higher interest rates, and so are more likely to be monitored. The interest rates $F^1$ and $F^2$ will of course depend on the type of interaction between the bank and the firm: long-term contract or long-term

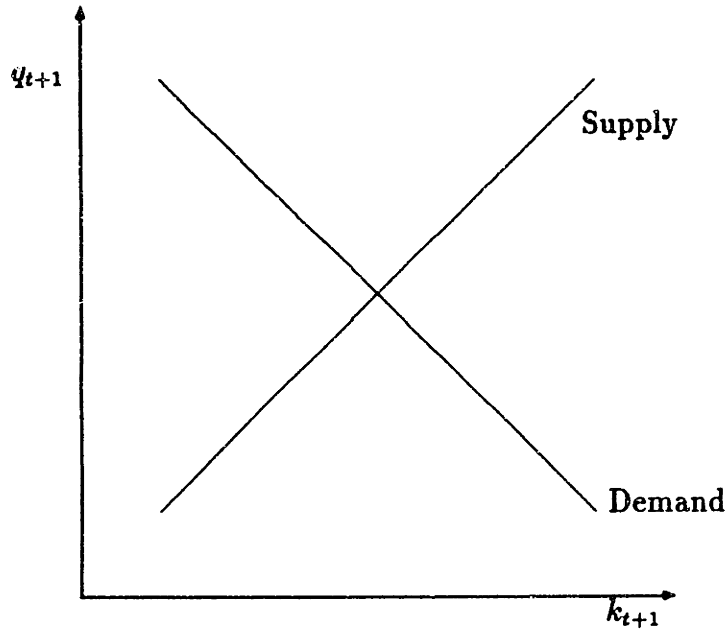



Figure 2-7: Equilibrium

relationship.

Per capita capital formation is

$$k_{t+1} = K\hat{\omega}_t - \int_0^{\hat{\omega}_t} (\gamma_1 + \gamma_2) G_K(F^1(\omega))\, d\omega \tag{2.12}$$

$$+K\hat{\omega}_{t-1} - \int_0^{\hat{\omega}_{t-1}} \gamma_1\, G_K(F^2(\omega))\, d\omega;$$

(2.12) is the capital supply function. Four of the variables on the right-hand side vary with $q_{t+1}$: $\hat{\omega}_t$ and $\hat{\omega}_{t-1}$ increase with $q_{t+1}$, because a higher price for capital will make more firms creditworthy; $F^1$ and $F^2$ decrease with $q_{t+1}$, because a bank needs to charge less if the capital with which it is repaid is worth more. Differentiate (2.12) with respect to $q_{t+1}$ to see that $dk_{t+1}/dq_{t+1} > 0$, as expected for a supply curve.

Demand for next period's capital stock is determined by setting its price equal to its marginal product:

$$q_{t+1} = f'(k_{t+1}) \tag{2.13}$$

Combining (2.12) and (2.13) yields Figure 2-7.

Within a period, equilibrium is determined by the intersection of capital supply and demand, as shown in Figure 2-7. However, the position of the capital supply curve depends on $q_{t+2}$, because the number of young firms able to obtain credit will change with the price of the capital they produce when old. In this sense the equilibrium of the model is forward-looking. The equilibrium is also backward-looking, as the number of old firms investing in period $t$ will depend on how many old firms got credit in period $t-1$, which depended on $q_t$. The equilibrium path of $q_t$ over time is the solution to the second-order non-linear difference equation produced by substituting (2.12) into (2.13):

$$q_{t+1} = f'\left(K\hat{\omega}_t - \int_0^{\hat{\omega}_t} (\gamma_1 + \gamma_2) G_K(F^1(\omega))\, d\omega \right. \tag{2.14}$$

$$\left. + K\hat{\omega}_{t-1} - \int_0^{\hat{\omega}_{t-1}} \gamma_1\, G_K(F^2(\omega))\, d\omega \right),$$

where $\hat{\omega}_t$, $\hat{\omega}_{t-1}$, $F^1$ and $F^2$ are implicitly functions of $q$ at different times.

One solution to (2.14) is a stationary solution where the price of capital and the capital stock are both constant over time. To see that a stationary solution exists and is unique, note that $dk/dq > 0$ along the capital supply curve and $dk/dq < 0$ along the capital demand curve.[19] In the stationary solution, a constant fraction of each generation of firms gets credit and invests. The variables $q$, $k$ and $\hat{\omega}$ do not vary over time; neither do the interest rates $F^1(\omega)$ and $F^2(\omega)$.

Any equilibrium path of $q_t$ that differs from the stationary solution will have the following feature: a time period with expensive, scarce capital will be followed by a period with cheap, abundant capital. To see this, look at Figure 2-8. The central capital supply curve $S_2$ represents the stationary equilibrium where $q_{t+2} = q_{t+1}$. Along the upper capital supply curve ($S_1$), the same price of capital in period $t+1$ brings forth less investment than in the stationary equilibrium; for this to be the case, $q_{t+2}$ must be lower than in the stationary equilibrium, meaning fewer firms can finance

[19] To prove $dk/dq > 0$, we must deal with the fact that $dF^1/dq_{t+2}$ will be positive for firms with $\omega$ just below $\omega^*$ because one effect of higher $q$ is tougher outside competition for the bank in the second period, lower second period profits, and a higher first period interest rate. For other levels of $\omega$ this effect will be dominated by the usual effect: higher $q$ means a lower interest rate is needed for the bank to cover its cost of funds. It is reasonable to assume the sum over $\omega$ of all these effects is negative so that $dk/dq > 0$.

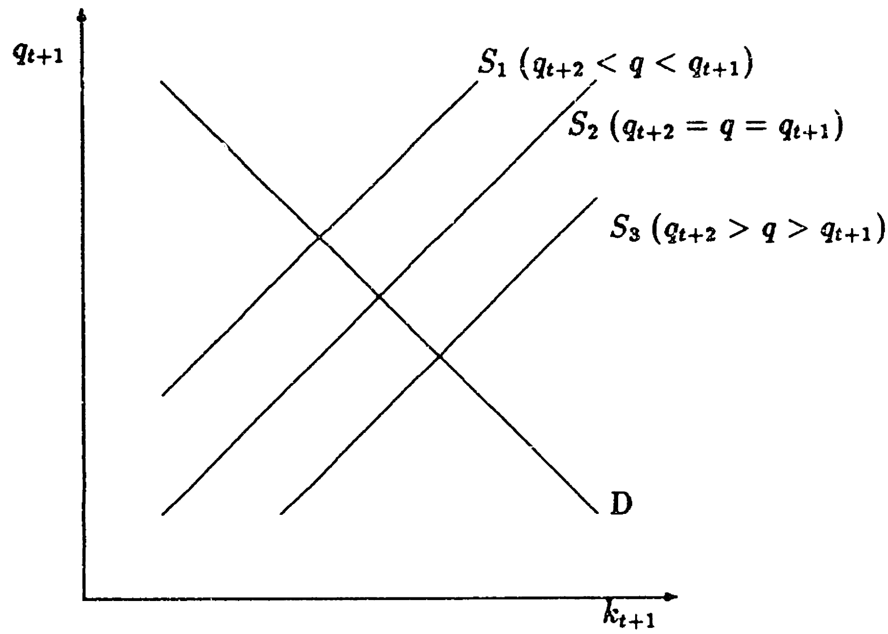



Figure 2-8: Oscillating equilibrium

investment. By the same logic, along $S_3$ $q_{t+2}$ must be higher than in the stationary equilibrium. Any equilibrium path of $q_t$ that differs from the stationary equilibrium must have $q_t$ alternating between high and low values. Anytime that $q_{t+1} > q$ (such as at the intersection of $S_1$ and $D$), $q_{t+2}$ must be less than $q$ and vice versa.

## 2.5 Extensions and conclusions

### 2.5.1 Rationing

Throughout the paper I have assumed the supply of savings is ample enough to fund every profitable investment project. I did so to ensure that storage was always positive, tying down the equilibrium interest rate at $R$ and making it easier to analyze the bank's lending decision. If this condition didn't hold and banks were unable to lend to all the firms whose projects are creditworthy, the banks would have to ration the available credit among firms.[20] Once rationing occurs, in exchange for the added complexity of worrying about the equilibrium interest rate, I can tell a richer story

[20]The term "rationing" as used here has a different meaning from "credit rationing" as defined by Keeton (1979) or Stiglitz and Weiss (1981).

of credit allocation that includes determining how banks allocate a limited supply of credit between prospective borrowers.

While I will not solve out the model with rationing, I will note several features of a rationing equilibrium. If rationing occurs and savings declines, a bank will always serve its old customers first, under both a long-term contract and a long-term relationship. Under a long-term contract, the bank will fulfill its obligations under existing contracts by supplying funds to old customers before it takes on new obligations by signing contracts with new customers.[21] Under a long-term relationship, old customers yield positive profits for the bank while new customers yield no profit; banks will always lend to old customers first. Credit allocation proceeds this way even though some young firms are more productive than some old firms. The bank's profit motive makes its interest diverge from the social good.

If a bank operating with long-term relationships had to sever ties with some old borrowers, it would keep longest those firms whose loans provide it with the largest second-period profit. Those will be the firms with average productivity, as Figure 2-6 makes clear. The first to be rationed would be the high productivity firms on whose loans the bank earns no profit (firms with $\omega \leq \underline{\omega}$). The bank does not even channel funds to the most productive of its old customers, because they are not the most profitable. Blackwell and Santomero (1982) came to the same conclusion for the same reasons, though in a very different model.

### 2.5.2 Bank failures

Return to the version of the model with no rationing and suppose that each period with some probability a single, randomly chosen bank fails because of fraud. When a bank fails, its loans are repaid, but the money vanishes and the depositors receive nothing. The terms of all loan contracts will change slightly, to reflect the chance that the lending bank will fail. Some savers lose their savings, but that has no effect on investment as credit is not rationed. More importantly, some firms will see their

[21] A rationing equilibrium would make the long-term contract more costly to write and therefore less likely to occur.

relationship with their bank disappear. These firms will have to turn to another bank for credit to invest in the second period of their life. Some of the firms that lost their banking relationship will be able to get a one-period loan as a new borrower; firms with $\omega \leq \omega^*$ will be able to get credit for just one period. However, less productive firms with $\omega \in (\omega^*, \hat{\omega})$ will not be able to replace the credit they were expecting from their failed lender, because their investments are not profitable enough to cover the higher cost of monitoring. They were only able to get credit in the first period because the bank took a loss, expecting to earn profits off the relationship in the second period. Aggregate investment will fall, temporarily, following a bank failure. It will return to its previous level once all firms who were affected by the failure have died.

### 2.5.3 More than three periods

If the firms and agents in this economy lived for more than three periods, the interest rates charged under a long-term relationship would come closer to those charged under a long-term contract. If the horizon were longer, outsider banks would provide stiffer competition for insider banks because they could reap profits from dealing with the firm for many future periods, instead of just one. More firms would fall into the range $(\underline{\omega}, \omega^*)$ and fewer would be in the range $(\omega^*, \hat{\omega})$. Fewer firms would be charged the profit-maximizing interest rate $\overline{F}(\gamma_1)$.

### 2.5.4 Conclusion

The model in this paper is an improvement upon previous research into the theoretical nature of bank lending because it explicitly considers the long-term interaction between borrowers and lenders and embeds it in a general equilibrium framework. Long-term relationships are an important feature of financial markets; a good model should feature them. Once a bank and a firm have repeated dealings, an immediate result is that more firms can get credit and invest.

Firms of different productivities are affected differently by long-term banking relationships. Very productive firms are unaffected, because publicly-available information is sufficient to induce savers to lend to them. Such lending would not have to go through a bank; borrowing by these firms is akin to a commercial paper market where a firm's creditworthiness alone backs up its debts. The least productive firms are not helped by a long-term banking relationship because a long-term relationship cannot reduce monitoring costs enough to allow these firms to get external credit. Some firms in this group may still have profitable investment opportunities, but only if they are financed with internal funds that carry no monitoring costs.

The middle-productivity firms are affected most by long-term banking relationships. More of these firms get credit once a long-term relationship is taken into account by the lender. For these firms to get credit, more information is required than what is publicly available. Banks produce that information cheaply by lending money under a debt contract. Identifying which firms are helped by long-term banking relationships is the main contribution of the model in this paper.

Whether or not a bank can commit not to exploit the market power it gains by making a loan to a young firm determines whether the bank provides funds under a long-term contract or a series of short-term contracts. Commitment, if possible, reduces the dead-weight loss from monitoring, making society better off.

The extension of the model to allow rationing in equilibrium is another place where explicitly considering a long-term relationship makes a difference. When rationing occurs, the bank continues to make loans to old customers, even though some more productive new customers may be denied credit. A shortfall of savings reduces the quantity of investment, but the interaction of a savings shortfall and long-term relationships reduces the efficiency of investment, because credit goes to some less productive firms.

# Chapter 3

# Bank Failures, Economic Activity, and the Great Depression

## 3.1 The role of bank failures

Can a disruption in the bank lending market worsen a business cycle? And was the Great Depression an example of a recession that was deepened by a banking collapse? Bernanke (1983) attempted to address these two very important macroeconomic questions, but for reasons that will be discussed below, he failed to provide decisive evidence. I construct two tests that can confirm or deny the link from bank failures to the real economy posited by Bernanke. Both tests suggest bank failures did worsen the Depression, but neither test is very powerful, weakening my conclusions.

On the question of whether disruptions in the supply of bank credit can worsen business cycles, macroeconomists are of three minds. First come those who believe bank lending always responds to and never affects economic activity. Real business cycle theorists make up one part of this group; they implicitly ascribe limited or no importance to bank credit by omitting it from their models entirely.[1] Those macroeconomists who do consider banking markets can be divided by the relative importance they assign to monetary and non-monetary effects. The "monetarists," led by Fried-

[1]One exception is Díaz-Giménez *et al* (1992).

man and Schwartz (1963), believe that bank failures play a role in macroeconomic fluctuations because bank deposits make up an important part of the money supply. The "non-monetarists," led by Bernanke (1983), hold that disruptions to the market for bank credit have their own negative effect on the economy, even without adversely affecting the money supply. Bank credit is important because banks are a special sort of financial intermediary. They provide credit to firms that can not raise funds in capital markets. The theoretical model in chapter 2 supports the "non-monetarist" argument; there I suggest that bank failures can have an adverse effect on output by disrupting established relationships between banks and firms.

The relative importance of the bank crisis among the various macroeconomic events of the Great Depression has been hotly debated for some time. Friedman and Schwartz (1963) discuss the banking crisis at length, giving the central role to the dramatic decline in the stock of money that resulted from it. Temin (1976 and 1989) denies the banking crises played an independent role; he blames the initial downturn in 1929–31 on a decline in spending and the severity of the Depression in 1931–32 on the deflationary pressure of the international gold standard. He views the banking crisis as an effect, rather than a cause, of the decline in output in 1929–33.

Bernanke (1983) advanced the thesis that, by disrupting the market for bank credit, the bank failures of 1930–33 had an important negative effect on output on top of the decline in the stock of money to which they contributed. The regression around which his argument centered was

$$y_t = \alpha(L)\, y_{t-1} + \beta(L)\, \tilde{m}_t + \gamma(L)\, DBANK_t + u_t \tag{3.1}$$

where $y$ is a measure of the growth rate of economic activity, $\tilde{m}$ is the unexpected movement in the money supply, $DBANK$ is the change in bank failures, normalized by the price level, and $L$ is the lag operator. He estimated (3.1) by ordinary least squares (OLS) and interpreted significantly negative coefficients on bank failures as evidence that bank failures caused declines in output over and above the declines they produced in the money supply.[2] However, bank failures could be a product of declines in economic activity, rather than a cause. Because of the potential reverse causation, a negative correlation between economic growth and bank failures of the kind demonstrated by Bernanke cannot by itself provide definitive evidence either for or against a causal role for bank failures. In this paper, I do two things to try to provide such evidence: first, I test for reverse causality by exploiting the variation in bank regulation across states during the Great Depression; second, I employ bank capital as an instrumental variable to consistently estimate the effect of bank failures on economic activity and provide a second test for the presence of reverse causality. The paper proceeds as follows. In section 3.2 I motivate both procedures at an intuitive level and establish the goal of each test. I lay out the implementation, data requirements and results of each test in turn in sections 3.3 and 3.4. A final section draws some conclusions.

## 3.2 Two approaches to the reverse causation problem

A negative coefficient on bank failures, when regressed on output, has been used to support the notion that financial distress can harm the real economy. Alternatively, the correlation between declines in economic activity and bank failures may only be due to the fact that when times are bad, a bank's assets fall in value and bank insolvency becomes more likely. Making this reverse causation argument requires some attention to the timing of the regression. If a decline in economic activity takes time to affect bank solvency, reverse causation is not a problem and the coefficient on bank failures in equation (3.1) is consistent. If a negative macroeconomic shock causes banks to fail within the same month, the coefficient on bank failures is inconsistent. No one could deny the general proposition that macroeconomic conditions affect bank solvency, but for Bernanke's regression coefficients to be consistent, all that is required

[2] See Table 2 in Bernanke (1983), p. 269.

is for aggregate shocks to affect the solvency of banks with a lag of one month or more.

Going the other way, I expect a bank failure to affect economic activity right away, since most bank credit to firms is short-term. In the monthly regressions in section 3.3, I look for a bank failure to have an effect on economic activity in the same month or the following month. In the quarterly regressions in section 3.4, I look for an effect in the same quarter only. Though I believe the response of economic activity to a bank failure will be quick, I want to avoid taking a stand on how the response to a bank failure is divided between economic activity in the current month and in the following month. For that reason, as I analyze the monthly regression results in section 3.3, I focus on the sum of the coefficients on current and lagged bank failures and their joint significance, and not on the magnitude of the individual coefficients. The simultaneity of bank failures and economic activity could be a problem for both current and lagged bank failures, since a bank could fail in advance of a decline in economic activity if the decline were forseeable and it made the value of the bank's assets fall enough that the bank became insolvent. If a bank fails in advance of or contemporaneously with a decline in economic activity, the coefficients in a regression such as (3.1) will be inconsistent; if a bank fails well after such a decline, the coefficients will be estimated consistently.

Keeping in mind the fairly quick effect of aggregate shocks on output needed for reverse causality to be important, I proceed in two steps: first, I use the variation in branch banking across states to test whether simultaneity is a problem. Next, I use bank capital as an instrument for bank failures to consistently estimate the effect of bank failures on economic activity and to test whether instrumental variables estimation yields different coefficients than ordinary least squares, as it would if reverse causality were important. The rest of this section describes the two tests.

During the Great Depression, some states allowed banks to operate branches while others did not. For example, California allowed statewide branching while Illinois allowed no branching at all. In a unit banking state like Illinois, the adverse effects of a local shock to economic activity would fall completely on local banks, banks dependent on a small geographic area for their business. If crops failed around Springfield, the

First National Bank of Springfield would be hard hit and might fail. In a state like California, where banks operated branches, a local shock would affect local branches of some banks, but the banks would be better able to weather the shock because they would have business spread out across several branches. A crop failure near Fresno would affect the Fresno branches of the large California banks, but because these banks were large and diversified, they would not be at a high risk of failing. In fact, Calomiris (1990) has shown that banks operating branches were more stable than unit banks in the interwar period.

Bank failures in branch banking states are less volatile than bank failures in unit banking states. The potential reverse causality from economic activity to bank failures will be weaker in states without branch banking, because in those states more bank failures will be unrelated to *statewide* movements in economic activity. Though they could be related to local movements in economic activity, local shocks will net out on the statewide level. In a sense, bank failures are more exogenous in states without branch banking. By estimating equation (3.1) separately for states with and without branch banking, I devise a test of whether reverse causality is important.

The second part of the paper employs bank capital as an instrumental variable for bank failures. To control for the possibility that bank failures are correlated with the contemporaneous shock to economic activity, some instrument is needed that is correlated with bank failures but uncorrelated with macroeconomic shocks. Bank capital should be negatively correlated with bank failures, because capital is a bank's buffer against adverse shocks. A bank with more capital can sustain greater losses before going bankrupt. Conversely, when two banks each suffer a loss of a given size, the bank with less capital will be more likely to fail, *ceteris paribus.* Two goals are accomplished by using bank capital as an instrumental variable. First, I can estimate the effect of bank failures on economic activity consistently, whether or not reverse causality is a problem. Second, by performing a Hausman test, comparing the instrumental variables estimate with the ordinary least squares estimate, I test for the presence of reverse causality. This second test complements the state-level test described in the preceding paragraph.

## 3.3 Regulation of branch banking varies across states

The test in this section exploits the variation in regulation of branch banking across states to test for the endogeneity of bank failures. The identifying assumption essential to this test is that states allowing branch banking had more stable banking systems than states prohibiting branch banking. The previous section of the paper justified this assumption and cited some evidence to back it up. In this section it is a maintained assumption throughout.

Consider, for expositional purposes, a simplified version of equation (3.1) relating the growth rate of economic activity to money surprises and bank failures:

$$y_t = \beta \tilde{m}_t + \gamma BF_t + u_t. \tag{3.2}$$

Simultaneity is a potential problem because bank failures in a given month may be correlated with shocks to economic activity in that month. Implicitly, a second equation such as the following completes the system:

$$BF_t = \alpha y_t + v_t \tag{3.3}$$

and $\alpha$ should be negative, since strong economic growth should make bank failures less likely to occur.

The system comprised of equations (3.2) and (3.3) applies equally well on the national and state levels. The coefficients $\beta$, $\gamma$, and $\alpha$ are structural parameters that will not vary across states, whether a state permits branch banking or not. This is so because a *statewide* shock to economic activity or bank failures, which is what the residuals in equations (3.2) and (3.3) represent, will have the same effect regardless of banking structure. The diversification within a state that is the distinguishing feature of branch banking will not matter for a statewide shock. A statewide movement in economic activity will have the same effect on statewide bank failures in both branch and non-branch states. (This statement would not be true if the regression were run on the level of a region within a state, because within-state diversification would help the banks in such a region withstand a local shock.) What will differ across states with and without branch banking is the variability of bank failures, or, equivalently, the percentage of the variation in bank failures that can be explained by the economic activity variable $y_t$. Bank failures will be more variable in non-branch states because of incomplete diversification by banks in those states. In terms of the econometric model, the variance of $v_t$ will be larger in non-branch banking states.

I show in Appendix D that the asymptotic bias of the ordinary least squares estimate of $\gamma$ in (3.2) is

$$\text{plim } \hat{\gamma} - \gamma = \frac{\alpha\sigma_u^2(1-\alpha\gamma)}{\alpha^2\sigma_u^2 + \sigma_v^2}. \tag{3.4}$$

Let $\hat{\gamma}^{(B)}$ represent the estimate of $\gamma$ using only data from branch banking states and for non-branch states let the estimate be denoted by $\hat{\gamma}^{(NB)}$. Take the difference of the probability limits of the two estimators, keeping in mind that all parameters except $\sigma_v^2$ are identical for the two groups:

$$\text{plim } [\hat{\gamma}^{(B)} - \hat{\gamma}^{(NB)}] = \alpha\sigma_u^2(1-\alpha\gamma)\left\{\frac{1}{\alpha^2\sigma_u^2 + \sigma_v^{2(B)}} - \frac{1}{\alpha^2\sigma_u^2 + \sigma_v^{2(NB)}}\right\}. \tag{3.5}$$

Under the maintained hypothesis that $\sigma_v^{2(B)} < \sigma_v^{2(NB)}$, the term in curly braces in (3.5) is positive. The term $(1-\alpha\gamma)$ will also be positive. Observe that a statement about the difference of the two estimators is also a statement about $\alpha$;

$$\text{plim } [\hat{\gamma}^{(B)} - \hat{\gamma}^{(NB)}] < 0 \quad \text{implies} \quad \alpha < 0$$
$$\text{plim } [\hat{\gamma}^{(B)} - \hat{\gamma}^{(NB)}] = 0 \quad \text{implies} \quad \alpha = 0.$$

A test of $\hat{\gamma}^{(B)} = \hat{\gamma}^{(NB)}$ is also a test of $\alpha = 0$. If $\alpha$ were zero, reverse causality would not be a problem. Conversely, a non-zero $\alpha$ would cause the OLS estimate of $\gamma$ to be inconsistent. The possibility that $\alpha < 0$ is equivalent to $\hat{\gamma}^{(B)} < \hat{\gamma}^{(NB)}$. In this section of the paper I will test the null hypothesis

$$H_0 : \hat{\gamma}^{(B)} = \hat{\gamma}^{(NB)}$$

against the alternative

$$H_1 : \hat{\gamma}^{(B)} < \hat{\gamma}^{(NB)}.$$

Before I can perform the state-level test, I need data on economic activity and bank failures by state. The best available data from this period on economic activity at the state level are the state-level employment growth figures in the *Monthly Labor Review*, published by the U.S. Bureau of Labor Statistics.[3] The Bureau of Labor Statistics reprinted the reports of all state Departments of Labor who kept statistics on employment. In 1931, twelve states provided such data, somewhat irregularly. From February 1932 onwards, the *Monthly Labor Review* provides monthly employment growth data for all states. Monthly state-level data on bank failures is available in the *Federal Reserve Bulletin* from January 1931 through February 1933. Appendix E presents the key state data in graphical form. Availability of the bank failure data determined the time frame for the test, and availability of employment statistics for a sufficient number of months determined which states would be included. The sample fortunately encompasses two key years of the Great Depression and most of the large states.

In this test, the economic activity variable $y$ is employment growth in the state and the bank failure variable is the percentage of banks (measured by deposits) in the state that failed during the month.[4] This measure of bank failures is more appealing than Bernanke's (the change in deposits of failed banks deflated by the wholesale price index), because the distress in bank lending markets caused by a bank failure should be related to the size of the market. Failure of a bank with $1,000,000 in deposits would be more disruptive if it represented 2% of total bank deposits than if it

[3]See Wallis (1989) for a more in-depth discussion of the BLS data.

[4]Data on total deposits by state is not available monthly. I normalized each month's bank failures by the 1934 level of deposits in the state, taken from *Banking and Monetary Statistics*, pp. 24–33.

represented 0.2%, but under Bernanke's measure of bank failures both would be given the same weight. The regressions are panel estimates, restricting the states within a group to have identical coefficients. An important maintained assumption is that credit was local during this period, a necessary assumption for bank failures within a state to affect economic activity within the same state. Blanchard and Katz (1992) have shown that states have had persistent differences in employment growth over time. The estimation technique allows for this by time-differencing the data to remove the state-specific effect. The lagged dependent variable $y_{t-1}$ is instrumented for.[5] I estimated equation (3.1) for the entire twelve-state sample and then split the sample into "branch" and "non-branch" states according to the percentage of loans and investments in the state held by banks operating branches, using the best available data on branches. See Table 3.1 for a breakdown of the sample into branch and non-branch states.

| State | Percentage of loans and investments in banks operating branches |
|---|---|
| California | 79% |
| Maryland | 38% |
| Massachusetts | 43% |
| Michigan | 59% |
| New Jersey | 45% |
| New York | 71% |
| Pennsylvania | 34% |
| Illinois | 0% |
| Iowa | 0% |
| Oklahoma | 0% |
| Texas | 0% |
| Wisconsin | 19% |

Source: *Federal Reserve Bulletin*, April 1930, p. 154.

Table 3.1: Extent of branch banking, December 31, 1929

Table 3.2 shows the results of the test.[6] The first column confirms Bernanke's result: bank failures have an overall negative effect on the growth rate of employment,

[5]The instruments are $y_{t-2}$ and $\tilde{m}_{t-4}$.

[6]The regressions in Table 3.2 were also run with price shocks replacing money shocks; the results are similar and are not reported to save space.

| Sample: | 12 states | 7 states with branches | 5 states without branches |
|---|---|---|---|
| Variable | | | |
| $y_{t-1}$ | -0.29 | -0.34 | -0.22 |
| | (0.19) | (0.35) | (0.17) |
| $\tilde{m}_t$ | -.26 | -.11 | -.39 |
| | (.23) | (.41) | (.24) |
| $\tilde{m}_{t-1}$ | .74 | 1.2 | .13 |
| | (.34) | (.58) | (.38) |
| $\tilde{m}_{t-2}$ | .17 | .35 | .029 |
| | (.33) | (.62) | (.36) |
| $\tilde{m}_{t-3}$ | .33 | .56 | .16 |
| | (.33) | (.53) | (.39) |
| Bank failures$_t$ | .17 | -.13 | .24 |
| | (.20) | ([illegible]) | (.17) |
| Bank failures$_{t-1}$ | -.33 | -.95 | -.33 |
| | (.22) | (.69) | (.19) |
| p-value of an F-test excluding bank failures | 0.033 | 0.34 | 0.0045 |
| N | 199 | 120 | 79 |
| S.E. | .031 | .035 | .024 |

Sample: February 1931 - February 1933

Table 3.2: State-level panel regressions

judging by the sum of the two coefficients, and the effect is statistically significant (p-value = 0.033).[7] The second column contains results for states that had substantial branch banking. The hypothesis that the true effect of bank failures is zero for these states cannot be rejected (p-value = 0.34), but the inability to reject is not because the coefficients are small but because their standard errors are large. The effects of money shocks are roughly twice as large in the branched states as in the entire sample, for which I have no explanation. For the states without branches, in the third column, the results look more like the results for the entire sample. Bank failures again have a significant effect on employment growth (p-value = 0.0045) and the overall effect is negative; the coefficients on bank failures are measured fairly precisely even though only five states are included in the non-branch group (as opposed to seven in the branch group).[8]

Earlier in this section, I proposed to test the null hypothesis that the coefficients on bank failures for the branch and non-branch states were the same. Given the identifying assumption that the variance of bank failures is greater in non-branch states, the only way they could be the same is if there is no reverse causality. If this is the case, then either estimate of the effect of bank failures on economic activity is consistent. If they differ, we expect from the alternative hypothesis proposed above that the coefficient from the regression including branch banking states should be less than the coefficient from the regression including non-branch banking states. This is essentially a test for sub-sample stability, or a Chow test. The result of the Chow test is

$$\begin{aligned} F(2,190) &= 0.39 \\ \mathrm{p-value} &= 0.68; \end{aligned}$$

the hypothesis that the coefficients are equal cannot be rejected. Because of the difference in the effect of a money shock across the two groups, I also did the Chow

[7] All tests of significance in this paper have a size of 5 percent.

[8] If the states on the boundary of the two groups (Pennsylvania and Wisconsin) are excluded, the results are very similar.

test allowing both money shocks and bank failures to have different coefficients, again testing to see whether the coefficients on bank failures were equal. This test also cannot reject the hypothesis that the two groups have the same response to bank failures.[9]

It is clear from Table 3.2 that the coefficients on bank failures differ in the direction suggested by the alternative hypothesis. The negative effect of bank failures is estimated to be stronger in branch banking states, as it would be if reverse causality were more severe in those states. However, the standard errors of the estimated coefficients are too large for the difference between the branch and non-branch estimates to be statistically significant. The large standard errors suggest that the test lacks power.

The tests in this section of the paper use employment growth unadjusted for seasonal variation as the dependent variable. Not enough data points were available for monthly dummies to be added to the regressions (as would be preferable). Adding monthly dummies is the best way to deal with seasonality of the dependent variable because it allows the effects of the model to be estimated jointly with the seasonal effects. However, doing nothing about seasonality leaves unanswered the criticism that seasonality is driving the results. To check whether this was the case, I replicated the regressions underlying Table 3.2 using seasonally adjusted employment growth as the dependent variable.[10] None of the results changed, though the precision of the estimated coefficients fell, predictably.

## 3.4 Bank capital as an instrument

Using bank capital as an instrument for bank failures is a conceptually straightforward two-stage least squares (2SLS) regression. The only issues to address are the timing of the observations on bank capital and the exact form of bank capital, levels or

[9] A Chow test to see if money shocks have different coefficients for the two groups of states also could not reject the null hypothesis of equality across groups. The p-value was 0.43, which is less than the p-value of the test for bank failures reported in the text. The relationship between money shocks and economic activity in branch and non-branch states is left as a project for future research.

[10] Adjusted by subtracting off monthly means.

differences, to relate to bank failures when running 2SLS. Data on bank capital is only available quarterly, so all regressions in this section of the paper will use quarterly data. Using quarterly data will worsen any reverse causality that exists, because banks that failed one or two months after a bad macroeconomic shock will now be counted as failing contemporaneously with the shock. For the present purposes, this should help because it will heighten the difference between the OLS and 2SLS estimates under the hypothesis that reverse causality is a problem. The timing of observations is important. A higher level of bank capital at the *beginning* of the quarter makes bank failures during the quarter less likely. During a quarter, aggregate bank capital may fall as banks fail, but that is not an economically meaningful relationship, nor would it make bank capital a useful instrument. In what follows the observations on bank capital are beginning-of-quarter figures.

The correct relationship between bank capital and bank failures to use in the 2SLS regression is a more difficult question to answer. Theoretically, a higher level of bank capital should reduce the level of bank failures. Figure 3-1 depicts the relationship between aggregate bank capital and bank failures.[11] Contrary to expectation, bank capital was high at the same time that bank failures were high, during the years 1930-33. The reasons are clear: banks accumulated capital during the boom years of the late 1920s. Once the Depression hit in 1929, capital fell but the level was still high relative to the beginning and end of the interwar era, which produces the positive correlation between the level of bank capital and bank failures seen in Figure 3-1. This spurious positive correlation seems to overwhelm the negative correlation of the theory, making the level of bank capital useless as an instrument for the level of bank failures.

Another potential relationship between bank capital and bank failures exists: the change in a bank's capital could be correlated with the probability that it fails. A fall in capital is an ominous sign for a bank; bad loans charged off could cause such a fall. If there is serial correlation in the shocks that hit a bank, a bank whose capital declines,

[11]Figure 3-1 covers only those banks that were members of the Federal Reserve System, since better capital data is available for those banks.

due to bad loans or poor management, will be more likely to subsequently fail. A study by Secrist (1938) verified this at the individual bank level: banks that failed had falling capital in the year before failure. This suggests that the first difference of bank capital would be a good instrument for the level of bank failures. Since bank capital is measured at the beginning of the quarter, using the first difference implies a drop in capital during a given quarter makes failure in the following quarter more likely. Figure 3-2 shows the relationship between the first difference of bank capital and the level of bank failures at the national level, again restricted to member banks. A negative relationship between bank capital and bank failures is visible in this figure. This relationship will be used in the 2SLS regression to follow. Because the bank failure variable in the regression is a first difference to match the first difference of economic activity on the left-hand side, the bank capital variable used as an instrument will be the second difference.

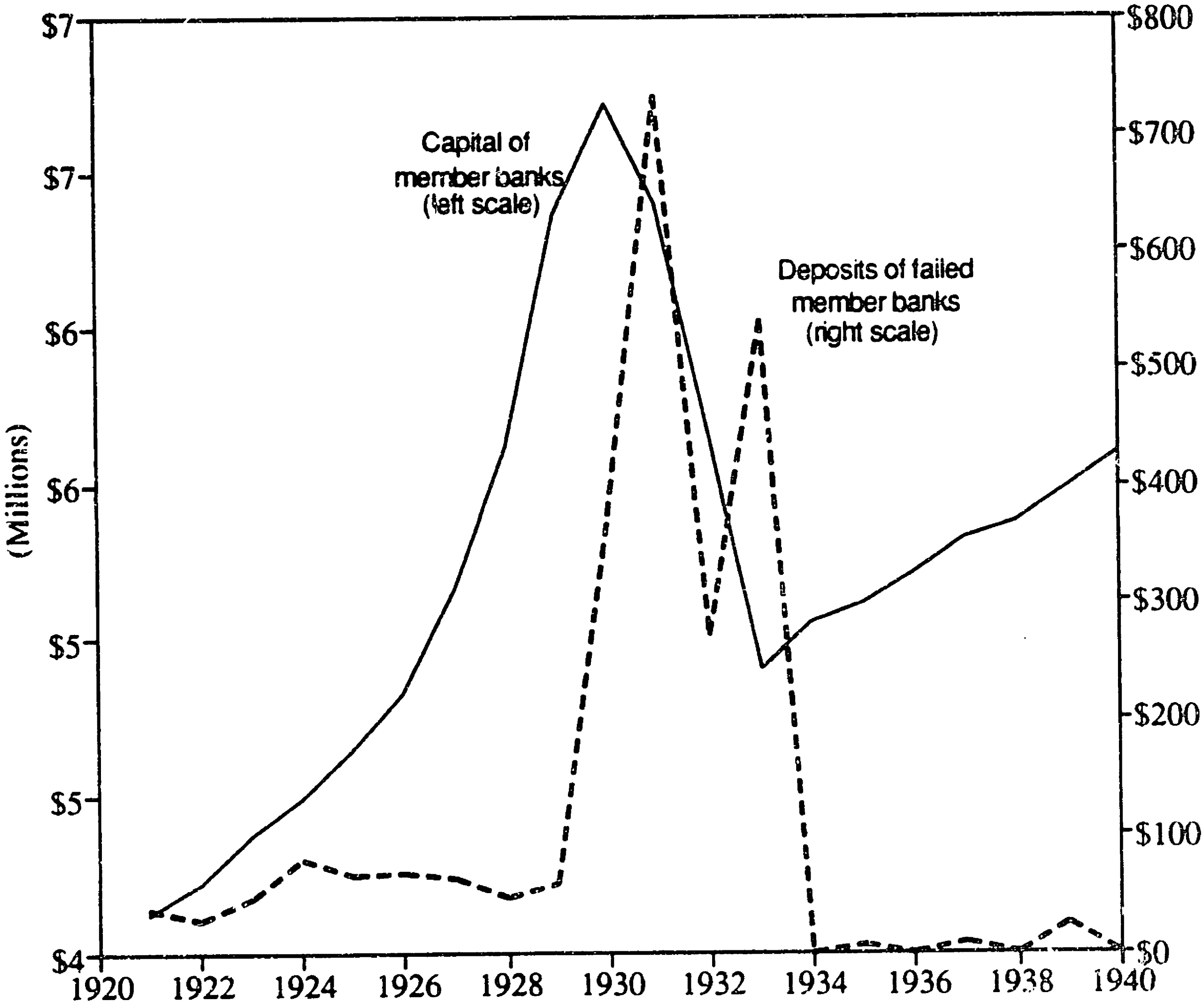
One might also think of using the ratio of the capital stock to some measure of bank size, like deposits or assets, as a proxy for the health of a bank. A graph with bank capital divided by deposits or assets would show the same positive correlation as Figure 3-1. A serious flaw in using such a ratio as an instrument is that movements in the ratio will be driven by movements in the denominator, because capital adjusts slowly relative to assets and deposits. Due to this problem, such a ratio will not be used.

The economic activity variable $y$ is here the growth rate of industrial production, not seasonally adjusted.[12] The unanticipated money shock $\tilde{m}$ (and the price shock $\tilde{p}$) is defined analogously to Bernanke (1983): the residual from the regression of money (price) growth on two lags of industrial production growth, money growth, and price inflation.[13] Bank failures are measured by the change in the percentage of banks (measured by deposits) that failed in a given quarter.[14] Appendix E presents the

[12]Industrial production is taken from *Industrial Production*, 1959 revision, Federal Reserve Board.

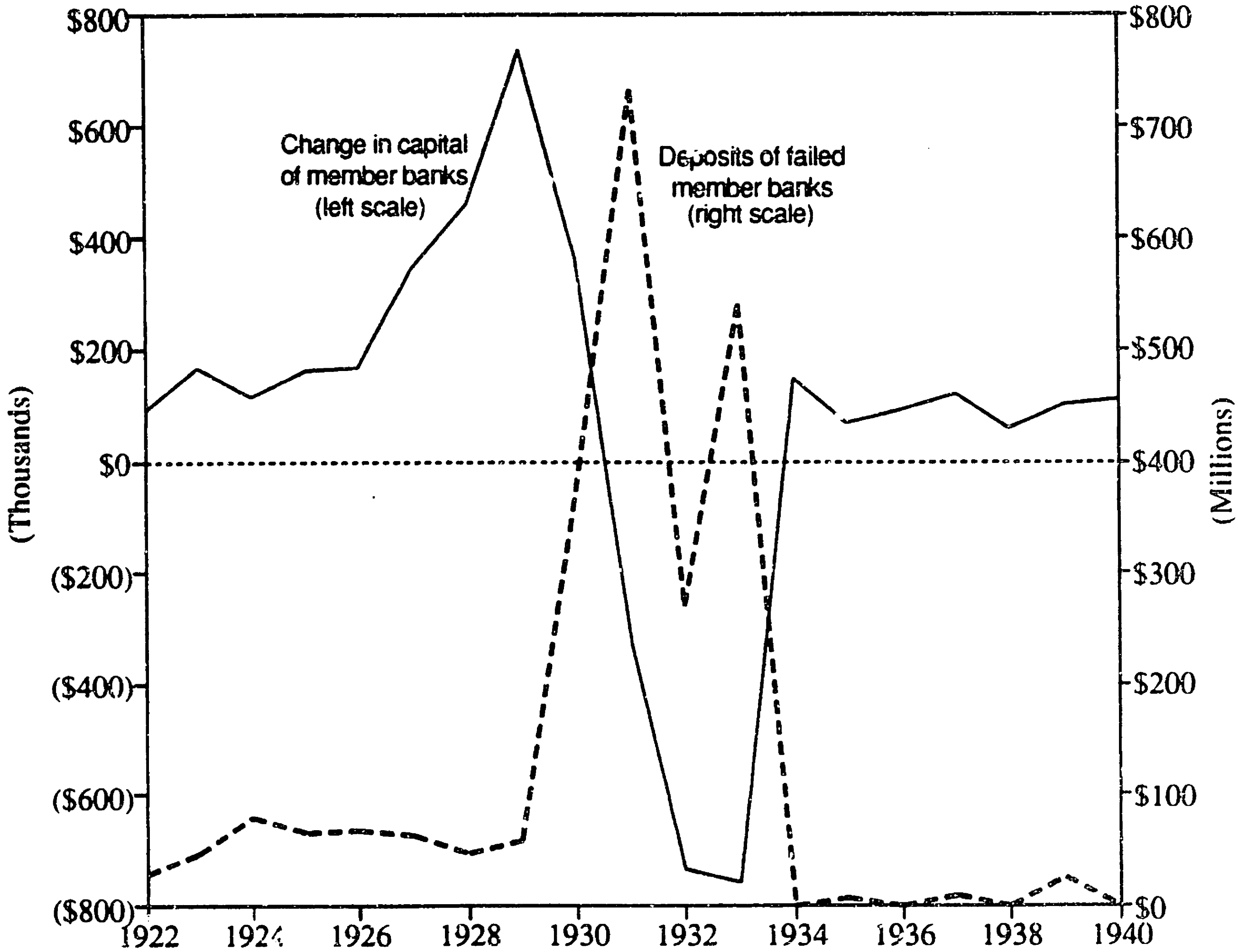
[13]The money supply data are from Friedman and Schwartz (1963), table A-1, column 7. The price level is the wholesale price index of the Bureau of Labor Statistics as reported in the *Federal Reserve Bulletin*.

[14]Bank failure data are taken from *Federal Reserve Bulletin*, September 1937 and subsequent issues. Deposit data are taken from *Banking and Monetary Statistics*. Following Bernanke, I adjust




Source: Banking and Monetary Statistics.
Bank failures for March 1933 have been multiplied by 15%. See text for explanation.

Figure 3-1: Capital and failures of member banks, 1921-1940




Source: Banking and Monetary Statistics.
Bank failures for March 1933 have been multiplied by 15%. See text for explanation.

Figure 3-2: Failures and the change in capital of member banks, 1922-1940 industrial production and bank failures series graphically.

The series on bank capital at the beginning of the quarter was based on call report data. The call reports, which Federal Reserve member banks submit three or four times per year, include the capital stock and surplus profits of all member banks. Because they are collected by a regulatory agency, these data are very accurate. The capital of all banks, including non-member banks, would be preferable but is not available. Member bank capital should move closely with all bank capital, since most large banks were members of the Federal Reserve System.[15] The capital series is the sum of all capital items except for reserves against bad loans; loan-loss reserves are excluded because, unlike regular capital, an increase in loan-loss reserves is not a sign of a healthy bank. Call reports were submitted three or four times a year; in those years when only three reports were submitted (seven of the twenty years in the sample), the missing quarter was interpolated.

Table 3.3 displays results of ordinary least squares and two-stage least squares regressions using bank capital as an instrument. The bank failure variable is the change in the percentage of total bank deposits in failed banks.[16] The first column of Table 3.3 shows an OLS regression that essentially replicates Bernanke's results. The sample period runs from the second quarter of 1921 to the fourth quarter of 1939; it is limited at the beginning by the availability of data on bank failures. Each regression included four quarterly dummy variables to capture seasonal movements in industrial production. The coefficients are not reported in the table to save space. The coefficient on lagged industrial production growth indicate it is stationary with a small amount of persistence. The monetary shock coefficients are positive, as expected. The bank failure variables are significantly negative, and large in magnitude. A major

---

the amount of bank failures in March 1933 downward by 85% to account for the bank holiday that resulted in many banks closing temporarily and then reopening after several months. The 85% number is arbitrary and is retained here for comparability with Bernanke's results.

[15]On June 30, 1930, deposits in member banks were $38 trillion and deposits in non-member banks were $13 trillion. In the years 1921–1940, member banks made up 50% of failed banks (weighted by deposits).

[16]For comparison with Bernanke's results, I ran both regressions using the change in real deposits in failed banks instead of the percentage of bank deposits in banks that failed; the results were similar to those in Table 3.3 and are available on request.

| Variable | OLS | 2SLS | OLS | 2SLS |
|---|---|---|---|---|
| $y_{t-1}$ | 0.27 | 0.30 | 0.24 | 0.35 |
| | (0.11) | (0.15) | (0.11) | (0.17) |
| $\tilde{m}_t$ | 0.99 | 0.69 | | |
| | (0.48) | (0.66) | | |
| $\tilde{m}_{t-1}$ | 0.97 | 1.6 | | |
| | (0.50) | (0.79) | | |
| $\tilde{p}_t$ | | | 1.5 | .89 |
| | | | (0.34) | (0.63) |
| $\tilde{p}_{t-1}$ | | | 0.60 | 0.12 |
| | | | (0.35) | (0.58) |
| Bank failures$_t$ | -6.4 | -23 | -3.2 | -17 |
| | (2.4) | (12) | (2.3) | (12) |
| N | 75 | 75 | 75 | 75 |
| S.E. | 0.068 | 0.089 | 0.063 | 0.080 |
| p-value of an F-test excluding bank failures | $< 0.01$ | 0.054 | 0.16 | 0.14 |
| p-value of the Hausman test | 0.052 | | 0.12 | |

Table 3.3: Ordinary and two-stage least squares results

episode of bank failures the size of the shock in the fourth quarter of 1930 would be associated with 7% lower growth in industrial production in the same quarter; the effect would die out quickly.

The 2SLS results in the second column of Table 3.3 closely mirror the OLS results.[17] Looking first at the predetermined variables (lagged industrial production and money and price surprises), the coefficients change slightly and the standard errors increase, as one would expect when moving from ordinary least squares to instrumental variables. Turning to the bank failure variables, the coefficients become more negative, which is not the direction they would move if the endogeneity of bank failures were a serious problem. If the instrumental variable technique were removing endogenous movements in bank failures from the regression, the coefficients would move closer to zero. The standard errors increase dramatically, highlighting the drawback of 2SLS estimation. The regression with price shocks replacing money shocks give similar results.

The proper way to test whether bank failures are endogenous is with a Hausman (1978) specification test. The Hausman test asks whether the difference between the coefficients on bank failures in the OLS and 2SLS regressions is statistically significant; if bank failures are endogenous, the difference should be negative. I implemented the Hausman test by projecting bank failures onto the instrument set, adding the projected variable to the regression and testing the hypothesis that those variables could be excluded. As the bottom line of Table 3.3 shows, a Hausman test comparing the OLS and 2SLS coefficients on bank failures shows that they are not significantly different. Even though the test comes close to the 5% critical level, it does not indicate that reverse causality is affecting the coefficients on bank failures, because the 2SLS estimates are more, not less negative. The Hausman test implies the original OLS estimates are consistent, but the large standard errors of the bank failure variables in the 2SLS regression suggest that the test is not a powerful one.

[17] In order for the printed standard errors to be correct, the residuals from these regressions must not be serially correlated. An LM test in the style of Godfrey (1988) showed that they are serially uncorrelated.

## 3.5 Conclusion

In this paper, I provide some further evidence on whether bank failures were a cause or an effect of the Great Depression. I conclude that the negative relationship between bank failures and economic activity cannot be wholly explained by saying banks fail because economic conditions worsen. The data support the conclusion of Bernanke (1983), that the collapse of the banking system was one of the main reasons the economy collapsed in 1929–33. The two tests I perform suffer from a lack of power, weakening the conclusions.

The results do not prove or even suggest that downturns in economic activity do not contribute to bank failures, merely that a shock to the macroeconomy does not trigger bank failures before the shock or within the same month. It is easy to believe there will be a lag of several months between a recessionary impulse and bank failures triggered by it. The results are consistent with a recession that begins in January causing banks to fail in April or May; the bank failures worsen an already bad situation by causing an additional decline in output. Of course, an independent shock could hit the banking sector and cause banks to fail, with output declines following.

The results of this paper contradict Temin (1976 and 1989) and other authors who deny bank failures a role in deepening the Great Depression. Though international linkages are surely important, the collapse of the U.S. banking sector was also a major contributor to the severity of the Great Depression in the United States. The questions posed in the introduction can now be answered in the affirmative. The banking collapse did play a large, independent role in deepening the Great Depression.

# Appendix A

# Mathematical Appendix to Chapter 1

## A.1 Derive the value of the bank's policy

The optimal policy for this type of a problem is of the form $(u, l)$, where $u$ and $l$ are the bounds where infinitesimal regulation takes place. Inside the bounds, $z(t)$ is allowed to float freely. A proof that such a policy is optimal when the exogenous variables follow geometric Brownian motion can be found in Bertola (1988), chapter 1.[1] To determine the parameters of the optimal policy in the present model, I find the value of an arbitrary $(u, l)$ policy and then solve for the optimal $u$ and $l$. My presentation follows Dumas (1989), extended to the case of geometric Brownian motion.

Between the barriers $u$ and $l$, $z(t)$ follows

$$dz(t) = \mu_z z(t)\,dt + \sigma_z z(t)\,dw_z. \tag{A.1}$$

At $z = u$, a costless regulator is applied. At $z = l$, a costly regulator is applied.[2]

[1] Bertola's proof also requires the assumption $f' > 0, f'' < 0$.

[2] A more formal presentation can be found in Harrison (1985), defining the regulated process $z(t)$ by:

$$\frac{z^{ur}(t)L(t)}{U(t)}$$

where $L$ and $U$ are the lower and upper regulator processes and $z^{ur}$ is the path $z$ would follow if

The value of an optimal program, given $u$ and $l$, is

$$V(z(t)) = E_t\left[\int_t^\infty e^{-\rho s} f(z(s))ds - dr\right]. \tag{A.2}$$

Assume $f(\cdot)$ is strictly concave and bounded. Rewrite (A.2) as a Bellman equation:

$$\begin{aligned} V(z(t)) \;&=\; \max\Big(E_t\left[f(z(t))dt + e^{-\rho dt}V(z(t+dt))\right], \\ &\quad \max_\eta (V(z+\eta) - \gamma 1(\eta > 0)\eta)\Big) \end{aligned} \tag{A.3}$$

$1(\cdot)$ is the indicator function. Remember that $0 \le z(t) \le 1$. Assume away the integer problem; the bank's control over $c(t)$ gives it perfect control over $z(t)$, allowing it to choose whatever increment or decrement $\eta$ it wants. When the bank chooses not to change the number of customers, (A.3) can be rewritten as

$$V(z(t)) = f(z(t))dt + (1 - \rho\, dt)\left[V(z(t)) + E_t\, dV(z(t))\right] \tag{A.4}$$

Using Ito's formula and (A.1),

$$dV = (\mu_z z V' + \frac{\sigma_z^2}{2} z^2 V'')dt + \sigma_z z V' dw_z \tag{A.5}$$

Take the expectation at time $t$ of (A.5), substitute into (A.4) and drop higher order terms to get

$$\rho V(z) = f(z) + \mu_z z V' + \frac{\sigma_z^2}{2} z^2 V''. \tag{A.6}$$

Since costs of adjustment are purely proportional, $V$ will never take a discrete jump. The following "value-matching" conditions must hold:

$$V'(u) \;=\; 0 \tag{A.7}$$

$$V'(l) \;=\; \gamma. \tag{A.8}$$

---

unregulated. $L$ and $U$ are increasing and continuous. $L(0) = U(0) = 1$. $U$ increases only when $z(t) = u$; $L$ increases when $z(t) = l$. I will follow the more intuitive presentation of Dumas, but the reader should keep in mind the rigorous foundation provided by Harrison.

These value-matching conditions ensure that the value function is continuous when an adjustment is made.

$V$ is the solution to the second order linear differential equation (A.6) with boundary conditions (A.7) and (A.8).[3] Let $\overline{V}$ be a particular solution to (A.6). ($\overline{V}$ will depend on the function $f(z)$.) The general solution is

$$\begin{aligned} V(z) &= \overline{V}(z) + K_1 z^{\delta_1} + K_2 z^{\delta_2} \qquad \text{(A.9)} \\ \text{where } \delta_1 &= \frac{1}{2}\left[ -\left(\frac{2\mu_z}{\sigma_z^2} - 1\right) + \sqrt{\left(\frac{2\mu_z}{\sigma_z^2} - 1\right)^2 + \frac{8\rho}{\sigma_z^2}} \right] \\ \delta_2 &= \frac{1}{2}\left[ -\left(\frac{2\mu_z}{\sigma_z^2} - 1\right) - \sqrt{\left(\frac{2\mu_z}{\sigma_z^2} - 1\right)^2 + \frac{8\rho}{\sigma_z^2}} \right]. \end{aligned}$$

If $\delta_1$ or $\delta_2$ is not a rational number, define $z^{\delta_1}$ as

$$z^{\delta_1} = e^{\delta_1 \ln z}.$$

The constants $K_1$ and $K_2$ are pinned down by the boundary conditions (A.7) and (A.8). Without specifying a functional form for $f(\cdot)$, (A.9) is the most explicit solution feasible.

## A.2 Choose $u$ and $l$ optimally

By the reasoning in the text, $u = 1$. At $z = l$, infinitesimal adjustment must be optimal. Put another way,

$$\begin{aligned} dz &= \arg\max_{\eta} \left(-\gamma\eta + V(l+\eta)\right) \\ \text{FOC} : 0 &= -\gamma + V'(l + dz) \\ 0 &= -\gamma + V'(l) + V''(l)dz \\ 0 &= V''(l) \qquad \text{(A.10)} \end{aligned}$$

[3] Dixit (1989) solves a similar equation.

and the last step uses (A.8). Equation (A.10) is the "smooth-pasting" condition for optimality. Since $V$ is known from (A.9), $l$ can be solved for.

## A.3 An explicit solution for a specific functional form

Assume

$$f(z) = z^{1/2} \; ; \; 0 \le z \le 1$$

which satisfies the conditions imposed on $f(\cdot)$ above $(f' > 0, f'' < 0)$. A particular solution to (A.6) is

$$\overline{V}(z) = \frac{z^{1/2}}{\rho - \mu_z/2 + \sigma_z^2/8}.$$

The boundary conditions (A.7) and (A.8) are

$$\begin{aligned} \frac{1/2}{\rho - \mu_z/2 + \sigma_z^2/8} u^{-1/2} + K_1\delta_1 u^{\delta_1 - 1} + K_2\delta_2 u^{\delta_2 - 1} &= 0, \\ \frac{1/2}{\rho - \mu_z/2 + \sigma_z^2/8} l^{-1/2} + K_1\delta_1 l^{\delta_1 - 1} + K_2\delta_2 l^{\delta_2 - 1} &= \gamma. \end{aligned}$$

These two equations determine $K_1$ and $K_2$ as functions of $u$ and $l$. Remember from above that $u = 1$. To find $l$, use the optimality condition (A.10):

$$\frac{-1/4}{\rho - \mu_z/2 + \sigma_z^2/8} l^{-3/2} + K_1(l)\delta_1(\delta_1 - 1)l^{\delta_1 - 2} + K_2(l)\delta_2(\delta_2 - 1)l^{\delta_2 - 2} = 0 \qquad \text{(A.11)}$$

where $K_1$ and $K_2$ are written explicitly as functions of $l$.

Equation (A.11) can in principle be solved for $l$ as a function of the model's parameters $(\mu_z, \sigma_z^2, \gamma, \rho)$, though an analytical solution is difficult. Numerical simulation provides the following comparative statics results:

$$\frac{\partial l}{\partial \mu_z} < 0; \; \frac{\partial l}{\partial \sigma_z^2} < 0; \; \frac{\partial l}{\partial \rho} < 0; \; \frac{\partial l}{\partial \gamma} < 0.$$

Each of these results is quite intuitive. When the drift of $z$ increases ($\mu_z \nearrow$), the bank can afford to wait longer to add customers because loans are more likely to increase without the bank paying the cost of adding a new customer. If the bank left $l$ unchanged when $\sigma_z^2$ increased, $z$ would hit the boundaries of its corridor much more often and the costs of adding new customers would mount. The bank will therefore widen its band by lowering $l$. The simulations show $l$ is relatively insensitive to $\rho$, but the partial derivative is negative. If the future becomes less important ($\rho \nearrow$), the bank will be less willing to pay $\gamma$ today in hopes of reaping profits from the customer in the future. The fourth result is obvious: if the cost of adjustment rises, adjustment will be less frequent.

# Appendix B

# Aggregation Appendix to Chapter 1

In this section of the appendix, I will assume that each bank regulates its loans/assets ratio within a band as the model suggests and derive the behavior of the aggregate ratio of loans to assets, $\bar{z}$. In what follows, assume $N$ equal-sized banks that follow identical $(l, u)$ band policies. The only remaining issue to be specified is how shocks to $z(t)$ are correlated across banks. First, I consider two relatively tractable polar cases: shocks are perfectly correlated across banks, and shocks are uncorrelated across banks. Then I consider a mixture of the two. Bertola and Caballero (1990) consider a similar question of aggregation, and I will loosely follow their methodology.

## B.1 Purely aggregate uncertainty

When shocks are perfectly correlated across banks, all uncertainty in the model is aggregate uncertainty. One aggregate shock affects the portfolios of all banks. An individual bank's $z_i(t)$ will follow:

$$\frac{dz_i(t)}{z_i(t)} = \mu_z dt + \sigma_z dw_z(t)$$

when no customers are added or dropped. Nothing on the right-hand side of the equation depends on $i$, so changes in $z_i(t)$ will be the same for all banks. The only difference in levels of $z_i(t)$ across banks will stem from differences in initial positions $z_i(0)$.

Given an initial distribution of $z_i(0)$ on $[l, u]$, it is straightforward to trace through the effects of a sequence of aggregate shocks. Figure B-1 does so. The top left graph displays an arbitrary cross-sectional density of $z_i(t)$; the height of the graph at any point $\zeta \in [l, u]$ is the number of banks with $z_i(0) = \zeta$. Over time, aggregate shocks move each bank's $z_i(t)$ around within the band $[l, u]$. Since all banks experience the same shocks, the cross-section distribution never changes, with one important exception. The exception occurs when a bank's $z_i(t)$ hits the edge of its band and customers are added or dropped. Such a bank's $z_i(t)$ will stay at the boundary while the rest of the $z_j(t)$'s move along with the aggregate shock. In the top panel of figure B-1, after several negative shocks the entire cross section density has shifted down and a mass has built up at $l$. After several positive shocks, a mass has built up at $u$ and the support of the distribution has shrunk further. As figure B-1 makes clear, as the cross-sectional distribution moves back and forth inside the band $[l, u]$, its support shrinks until it finally converges to a point distribution and all banks share the same $z(t)$.

Caplin and Leahy (1991) consider a similar case of purely aggregate uncertainty. In their model (which has nothing to do with banking), when a decision-making unit hits the boundary of its band it adjusts by an amount $s$. When the width of the band is $2s$, each unit that adjusts moves to the center of the band. The empirical cross section distribution, uniform with length $s$, will be invariant to the aggregate shock and will move around within the band of width $2s$ according to the aggregate shock. In my model, each bank adjusts by an infinitesimal amount when it hits the boundary of its band. The long-run empirical cross section distribution of banks in my model is uniform on an infinitesimal support, or simply a point distribution. This point mass moves around on $[l, u]$ following the aggregate shock.

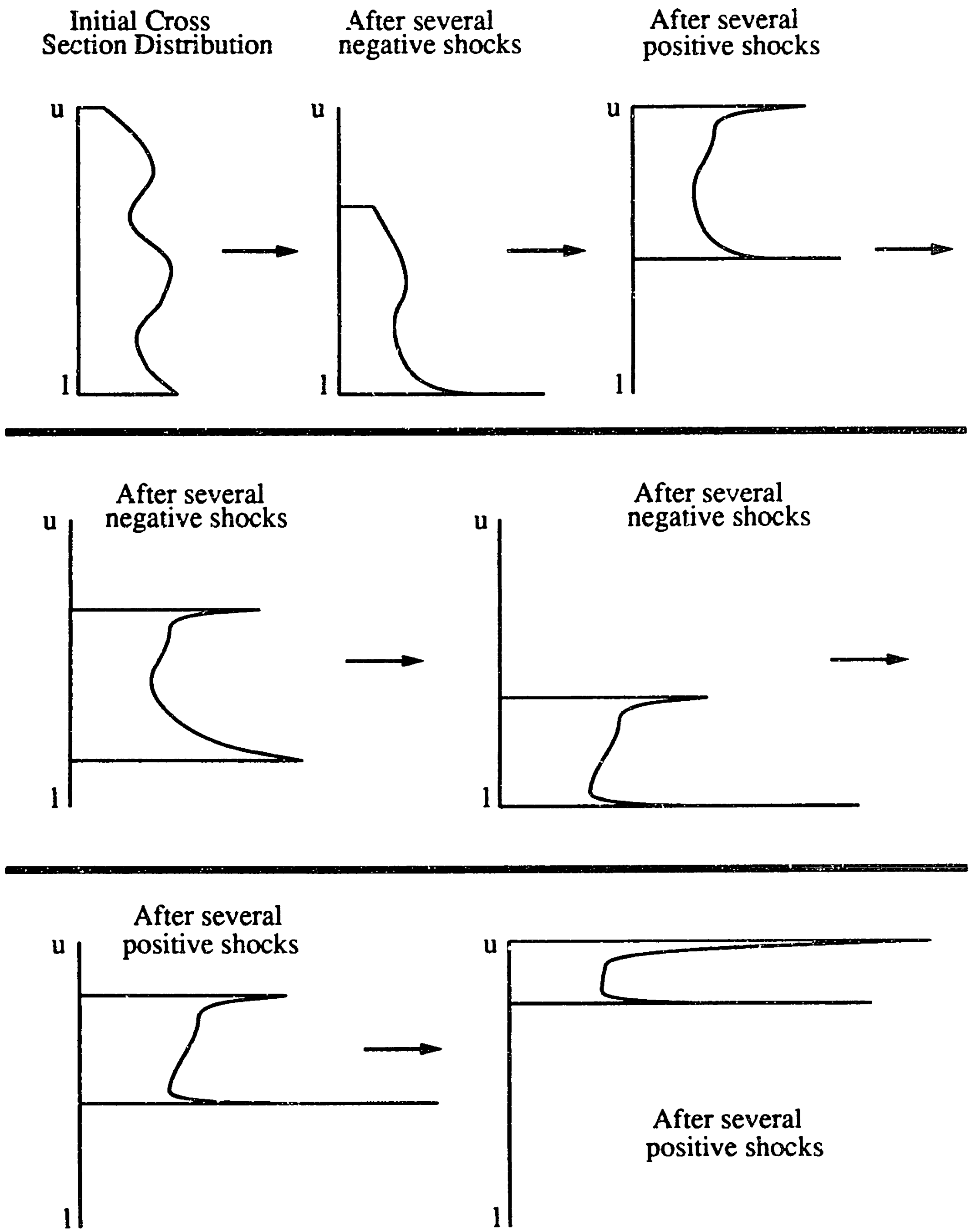



Figure B-1: Evolution of the cross sectional density — aggregate uncertainty

The aggregate ratio of loans to assets, $\bar{z}(t)$, is

$$\bar{z}(t) = \sum_{i=1}^{N} a_i(t) z_i(t) \tag{B.1}$$

where $a_i(t)$ is the share of bank $i$'s assets in total assets; $a_i = A_i / \sum_{i=1}^{N} A_i$. Since I have assumed all banks are the same size, $a_i = 1/N$. Even without that assumption, after the cross section distribution of $z_i(t)$ has converged to a point, so $z_i(t) = z(t)$ for all $i$, (B.1) can be written

$$\bar{z}(t) = z(t) \sum_{i=1}^{N} a_i(t) = z(t)$$

since $\sum_{i=1}^{N} a_i = 1$ by definition. When all uncertainty is aggregate, the correct dynamics for interpreting aggregate data are the dynamics for a single bank described in section 1.3.

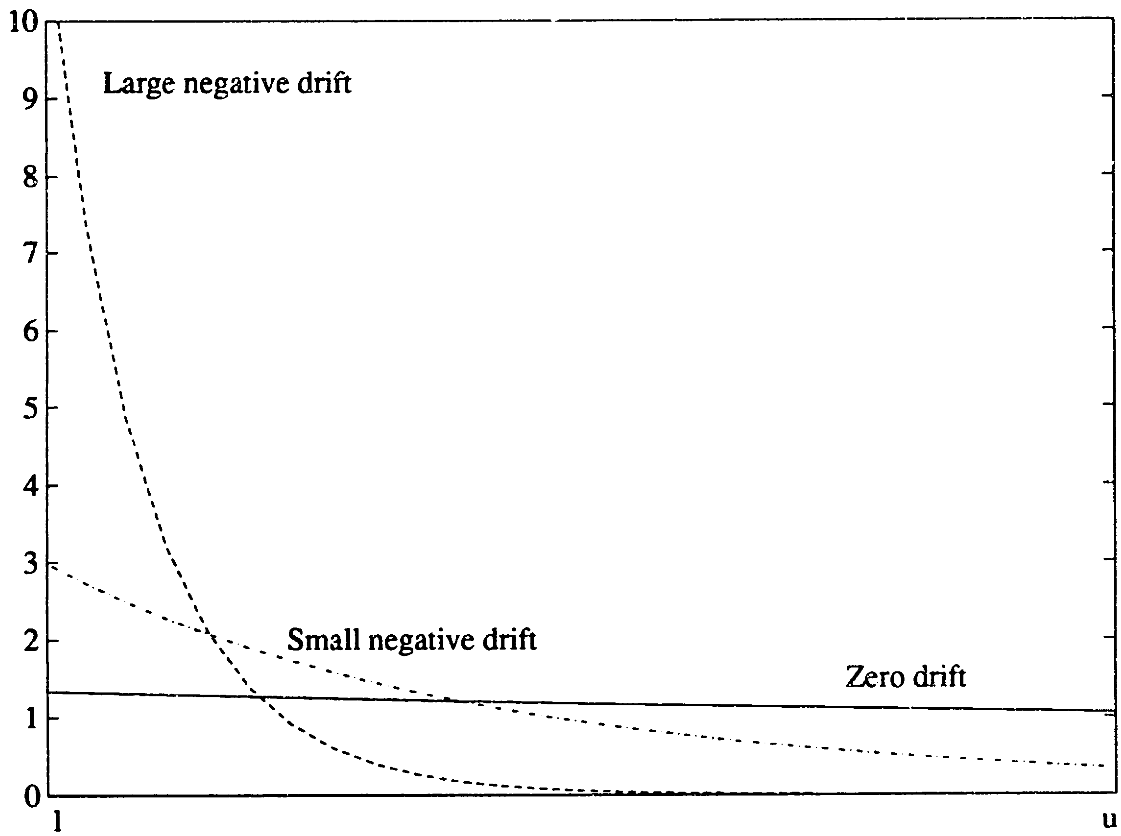
## B.2 Purely idiosyncratic uncertainty

When shocks are uncorrelated across banks, each bank faces purely local uncertainty and the position of one bank in the band $[l, u]$ is independent of the position of the other banks. In this case the empirical cross-sectional distribution of $z_i(t)$ on $[l, u]$ will not converge to a point distribution, as above; it will converge to the unconditional distribution of one bank on $[l, u]$.[1] The unconditional distribution of a single bank will be the ergodic distribution of $N$ banks. In section B.4 I show how to compute the unconditional distribution of a single bank on $[l, u]$ as a function of the drift and variance of $z_i(t)$. Figure B-2 shows the distribution for zero drift, small negative drift, and large negative drift. For zero drift, the distribution is flat. When the drift becomes negative, the distribution is skewed toward $l$. As the drift becomes more negative, the distribution is more skewed. (For positive drift simply think of the distributions in figure B-2 reversed, with the peak at $u$.)

Starting from an initial, arbitrary cross-sectional distribution of $z_i(t)$ on $[l, u]$, the

[1] If $N$ is large. See Bertola and Caballero (1990, p. 257).

Figure B-2: Unconditional distribution of $z_i(t)$



idiosyncratic uncertainty will spread banks out until their cross-sectional distribution matches the unconditional distribution of a single bank. The aggregate ratio of loans to assets, $\bar{z}(t)$ will be constant and will equal the mean of the distribution. From figure B-2 we see that $\bar{z}$ will depend on the drift of the process $z(t)$.

## B.3 Aggregate and idiosyncratic uncertainty

In reality both aggregate and idiosyncratic uncertainty exist. As I have just shown, aggregate uncertainty pushes all banks together into a point distribution, while idiosyncratic uncertainty spreads banks out according to the unconditional distribution of a single bank. With aggregate uncertainty, the aggregate variable $\bar{z}(t)$ follows the same dynamics as $z_i(t)$ for an individual bank. With idiosyncratic uncertainty, $\bar{z}(t)$ is constant. When both types of uncertainty are present, neither of the special cases outlined above will hold exactly. The cross sectional distribution will be neither a point mass nor invariant over time.

## B.4 Derive the probability density for each bank on $[l, u]$

When all uncertainty is idiosyncratic, as in section B.2, each bank's ratio of loans to assets follows

$$\frac{dz_i(t)}{z_i(t)} = \mu_z dt + \sigma_z dw_i(t)$$

where $dw_i(t)$ and $dw_j(t)$ are independent if $i \neq j$. Consider $\ln z_i(t)$, which follows an arithmetic Brownian motion,

$$d \ln z_i(t) = \mu_z dt + \sigma_z dw_i(t);$$

$\ln z_i(t)$ takes on values in the interval $[\ln l, \ln u]$. I represent the interval $[\ln l, \ln u]$ in discrete form as the $k \times 1$ vector $(\ln l, \ln l + \epsilon, \ln l + 2\epsilon, \cdots, \ln u - \epsilon, \ln u)$. The probability density of bank $i$ on $[\ln l, \ln u]$, $f_{it}$, can be represented as a $1 \times k$ vector

$(f_{it}(\ln l), f_{it}(\ln l+\epsilon), \cdots, f_{it}(\ln u-\epsilon), f_{it}(\ln u))$.

A discrete representation of the process for $\ln z_i(t)$ is

$$\Delta \ln z_i(t) = \begin{cases} +\epsilon & \text{with probability } 1-p = \frac{1}{2}(1+\frac{\mu_z dt}{\epsilon}) \\ -\epsilon & \text{with probability } p = \frac{1}{2}(1-\frac{\mu_z dt}{\epsilon}) \end{cases} \tag{B.2}$$

If $\epsilon = \sigma\sqrt{dt}$, as $dt \to 0$ the above process converges to the process for $z(t)$ given as (1.2) in the text.

This specification of $\ln z_i(t)$ implies the following Markov transition matrix, $P$, for $\ln z_i(t)$:

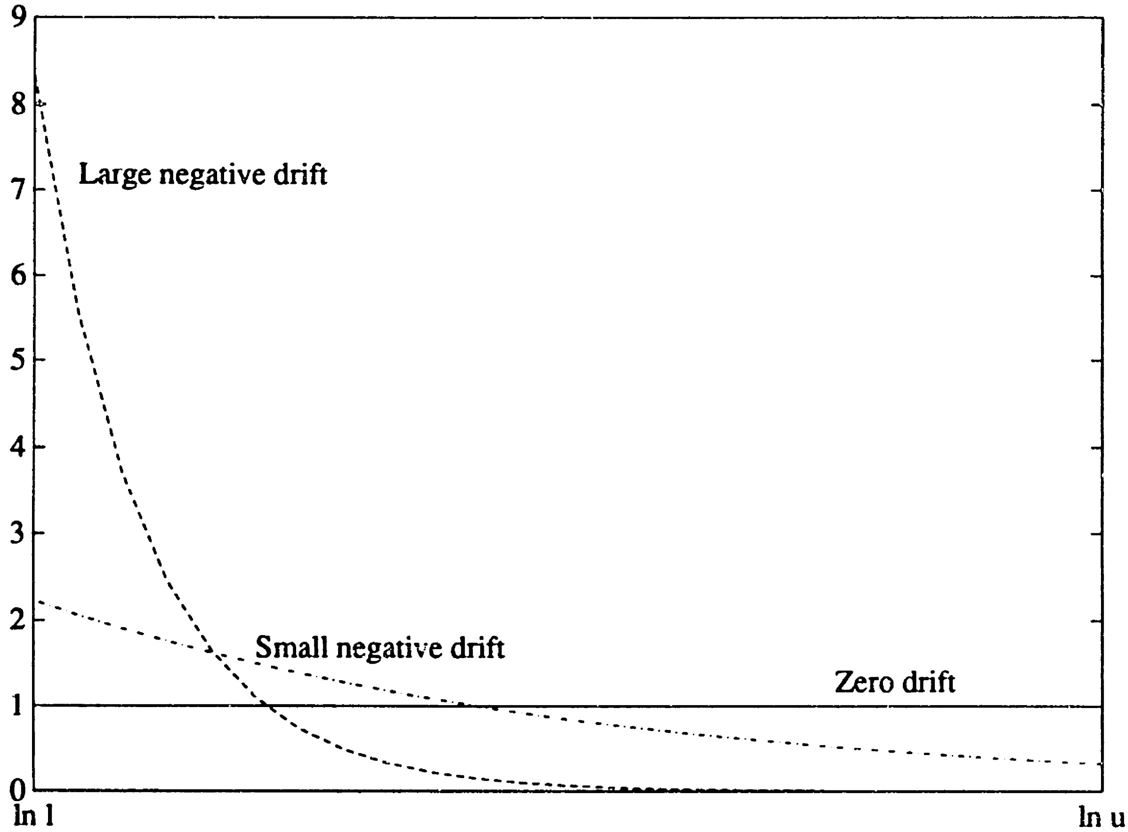
$$t \quad \begin{array}{r|ccccccc} & \multicolumn{7}{c}{t+dt} \\ \hline & \ln l & \ln l+\epsilon & \ln l+2\epsilon & \cdots & \ln u-2\epsilon & \ln u-\epsilon & \ln u \\ \ln l & p & 1-p & 0 & \cdots & 0 & 0 & 0 \\ \ln l+\epsilon & p & 0 & 1-p & \cdots & 0 & 0 & 0 \\ \ln l+2\epsilon & 0 & p & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \ln u-2\epsilon & 0 & 0 & 0 & \cdots & 0 & 1-p & 0 \\ \ln u-\epsilon & 0 & 0 & 0 & \cdots & p & 0 & 1-p \\ \ln u & 0 & 0 & 0 & \cdots & 0 & p & 1-p \end{array} \equiv P$$

The probability density of bank $i$ at time $t+dt$ is simply $f_{it+dt} = f_{it}P$. $f_{it}$ will converge to an ergodic (invariant over time) distribution $f$, where $f = fP$. Figure B-3 shows the ergodic distribution of $\ln z_i$ for different values of $p$.[2]

A larger $p$ implies a more negative drift $\mu_z$, as (B.2) makes clear. The three densities in figure B-3 are therefore labeled "Zero drift" for $p = .5$, "small negative drift" for $p = .52$, and "large negative drift" for $p = .6$. If the density of $\ln z_i$ is $f(\zeta)$ as plotted in figure B-3, the density of $z_i$ is $\frac{1}{\zeta}f(\zeta)$, which is plotted on page 109. In that graph $l = .75$ and $u = 1$.

[2]I generated the ergodic distribution $f$ by starting with an arbitrary initial distribution $f_0$ and iterating $f_{t+1} = f_t P$ until $f$ converged.

Figure B-3: Unconditional density of $\ln z_i(t)$



# Appendix C

# Data Appendix to Chapter 1

The series for $z(t)$ presented in figure 1-6 was constructed from Table 16, "Commercial Bank Assets and Liabilities—Last Wednesday-of-month series" in various issues of *Annual Statistical Digest*, Table 1.25 in the *Federal Reserve Bulletin* and Table 1.3 in *Banking and Monetary Statistics 1941-1970*. The data on $z(t)$ for national banks over the years 1863-1941 in figure 1-5 comes from Table 4, *Banking and Monetary Statistics*, p. 20.

The NBER peaks and troughs in table 1 were taken from Table 10, p. 178 of *Handbook of Cyclical Indicators* (1984).

The series in figure 1-1 were constructed from the quarterly survey on bank lending practices taken by the Federal Reserve Board. The Fed began reporting this survey with February 1967 (in the April 1968 *Federal Reserve Bulletin*). The survey was initially referred to as "Changes in Bank Lending Practices at Selected Large Banks." By 1981 the title was "Senior Loan Officer Opinion Survey on Bank Lending Practices, Selected Large U.S. Banks." Bankers were asked whether their practice concerning review of credit lines or loan applications had gotten firmer or easier over the last three months. They answered separately for established customers and for new customers. They could give one of five responses: much stronger, moderately stronger, essentially unchanged, moderately weaker, or much weaker. The number of banks giving each response was reported in the survey results. The series in figure 1-1 are weighted averages

$$\frac{\sum_{i=1}^{5} w_i n_i}{\sum_{i=1}^{5} n_i}$$

where $w_i$ is the weight for response $i$ and $n_i$ is the number of banks giving response $i$. The weights $w_i$ are given in table C.1.

| $i$ | Response | $w_i$ |
|---|---|---|
| 1 | Much firmer | -2 |
| 2 | Moderately firmer | -1 |
| 3 | Essentially unchanged | 0 |
| 4 | Moderately easier | 1 |
| 5 | Much easier | 2 |

Table C.1: Weights used to measure ease of credit

The data on loans made under commitment shown in figure 1-7 comes from Federal Reserve statistical release G.21. From July 1973 to June 1987, monthly, the Fed asked a group of large commercial banks about the size of their loans made under commitment and unused commitments. Paul Wachtel provided me with a complete time series of the G.21 data, adjusted by the Fed for breaks in the series.

# Appendix D

# Appendix to Chapter 3

Equation (3.4) in section 3.3 shows the asymptotic bias of the ordinary least squares estimator for $\gamma$ in equation (3.2). This appendix derives the asymptotic bias. The reduced form for the two-equation system (3.2) and (3.3) is

$$\begin{aligned} y_t &= \theta_1 \tilde{m}_t + \eta_{1t} \\ BF_t &= \theta_2 \tilde{m}_t + \eta_{2t} \end{aligned} \qquad \text{(D.1)}$$

Let $\hat{\delta} = [\hat{\beta}\ \hat{\gamma}]'$ be the OLS estimates from (3.2). By definition,

$$\hat{\delta} = \left[ \begin{pmatrix} \tilde{m}' \\ BF' \end{pmatrix} (\tilde{m}\ BF) \right]^{-1} \begin{pmatrix} \tilde{m}' \\ BF' \end{pmatrix} y$$

where $\tilde{m}$, $BF$, and $y$ are $T \times 1$ vectors. Using the reduced form (D.1),

$$\hat{\delta} = \left[ \begin{pmatrix} \tilde{m}' \\ (\theta_2 \tilde{m} + \eta_2)' \end{pmatrix} (\tilde{m}\ (\theta_2 \tilde{m} + \eta_2)) \right]^{-1} \begin{pmatrix} \tilde{m}' \\ (\theta_2 \tilde{m} + \eta_2)' \end{pmatrix} (\theta_1 \tilde{m} + \eta_1)$$

Take probability limits and drop out terms that go to zero to get

$$\text{plim } \hat{\delta} = \text{plim } \frac{1}{T} \begin{pmatrix} \tilde{m}'\tilde{m} & \theta_2 \tilde{m}'\tilde{m} \\ \theta_2 \tilde{m}'\tilde{m} & \theta_2^2 \tilde{m}'\tilde{m} + \eta_2'\eta_2 \end{pmatrix}^{-1} \text{plim } \frac{1}{T} \begin{pmatrix} \theta_1 \tilde{m}'\tilde{m} \\ \theta_1\theta_2 \tilde{m}'\tilde{m} + \eta_2'\eta_1 \end{pmatrix}$$

Assume

$$\text{plim } \frac{1}{T}\tilde{m}'\tilde{m} = M.$$

Then,

$$\text{plim } \hat{\delta} = \begin{pmatrix} M & \theta_2 M \\ \theta_2 M & \theta_2^2 M + \sigma_{\eta_2}^2 \end{pmatrix}^{-1} \begin{pmatrix} \theta_1 M \\ \theta_1\theta_2 M + \text{cov}(\eta_1, \eta_2) \end{pmatrix} \tag{D.2}$$

The reduced form variables $\theta_1$, $\theta_2$, $\eta_1$, and $\eta_2$ are related to the structural parameters by the following relationships:

$$\begin{aligned}
\theta_1 &= \frac{\beta}{1-\alpha\gamma} \\
\theta_2 &= \alpha\theta_1 \\
\eta_{1t} &= \frac{\gamma}{1-\alpha\gamma}v_t + \frac{1}{1-\alpha\gamma}u_t \\
\eta_{2t} &= \frac{\alpha}{1-\alpha\gamma}u_t + \frac{1}{1-\alpha\gamma}v_t
\end{aligned}$$

Evaluating the inverse in (D.2),

$$\text{plim } \hat{\delta} = \frac{1}{\sigma_{\eta_2}^2 M} \begin{pmatrix} \theta_2^2 M + \sigma_{\eta_2}^2 & -\theta_2 M \\ -\theta_2 M & M \end{pmatrix} \begin{pmatrix} \theta_1 M \\ \theta_1\theta_2 M + \text{cov}(\eta_1, \eta_2) \end{pmatrix}$$

Concentrating on the coefficient on bank failures ($\gamma$),

$$\text{plim } \hat{\gamma} = \frac{\text{cov}(\eta_1, \eta_2)}{\text{var}(\eta_2)} = \frac{\gamma\sigma_v^2 + \alpha\sigma_u^2}{\alpha^2\sigma_u^2 + \sigma_v^2}$$

As a comforting check on the algebra, note that if $\alpha$ is zero (so no simultaneity problem exists), $\hat{\gamma}$ is consistent for $\gamma$.

To find the asymptotic bias, subtract $\gamma$ from both sides to get

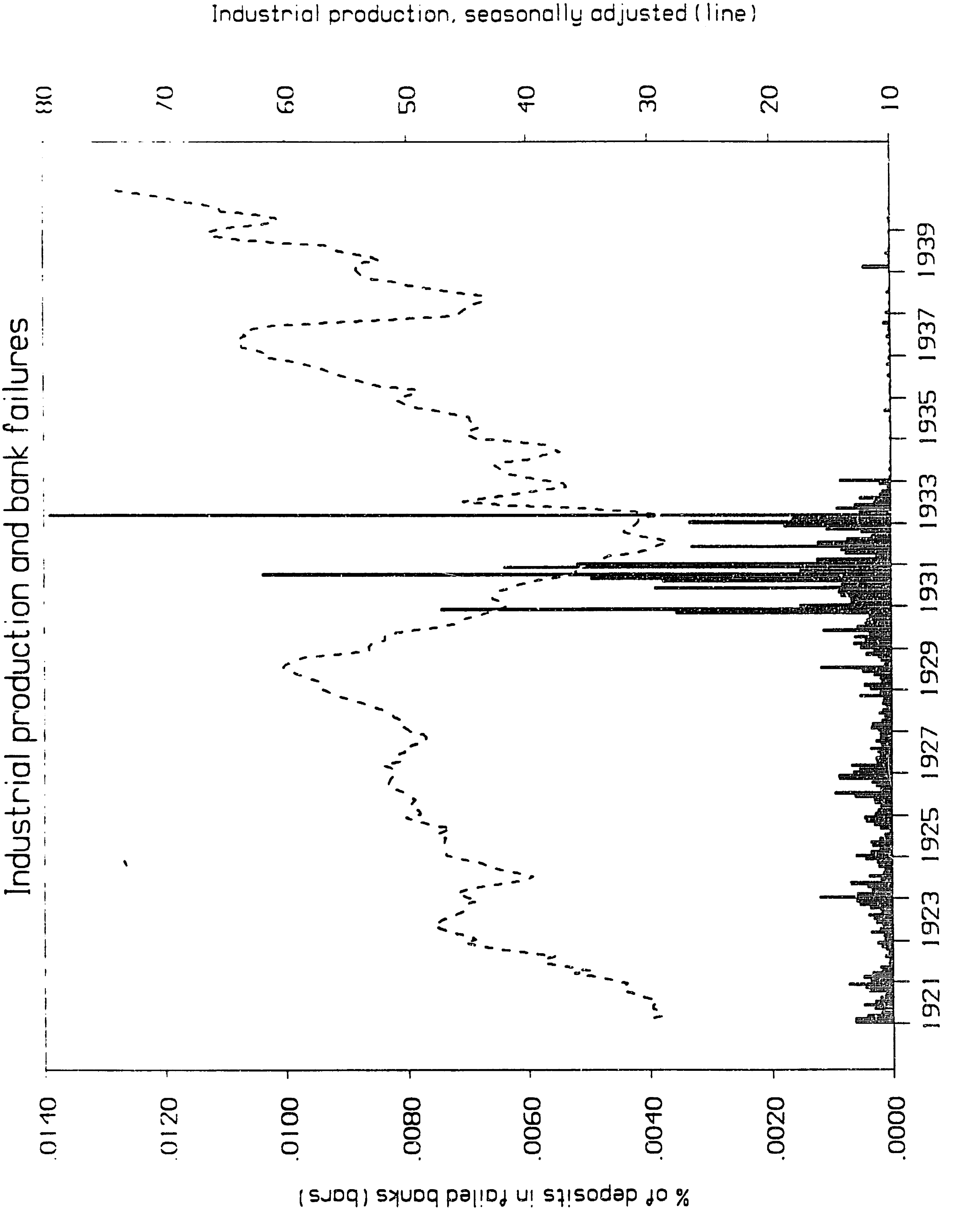
$$\text{plim } \hat{\gamma} - \gamma = \frac{\alpha\sigma_u^2(1-\alpha\gamma)}{\alpha^2\sigma_u^2 + \sigma_v^2}$$

which is equation (3.4) in section 3.3 of the text.

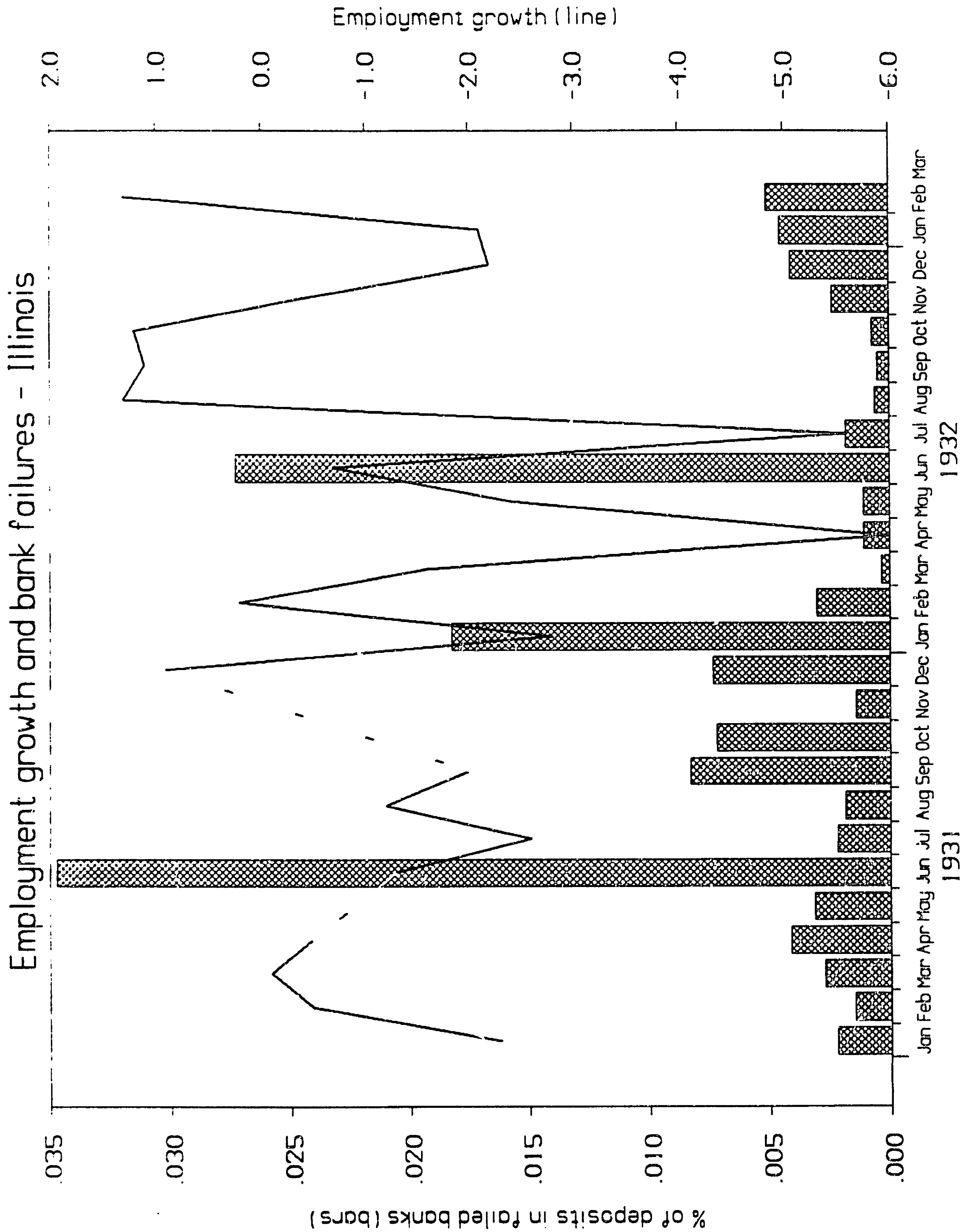
# Appendix E

# Data appendix to Chapter 3

Thirteen pages of graphs follow.

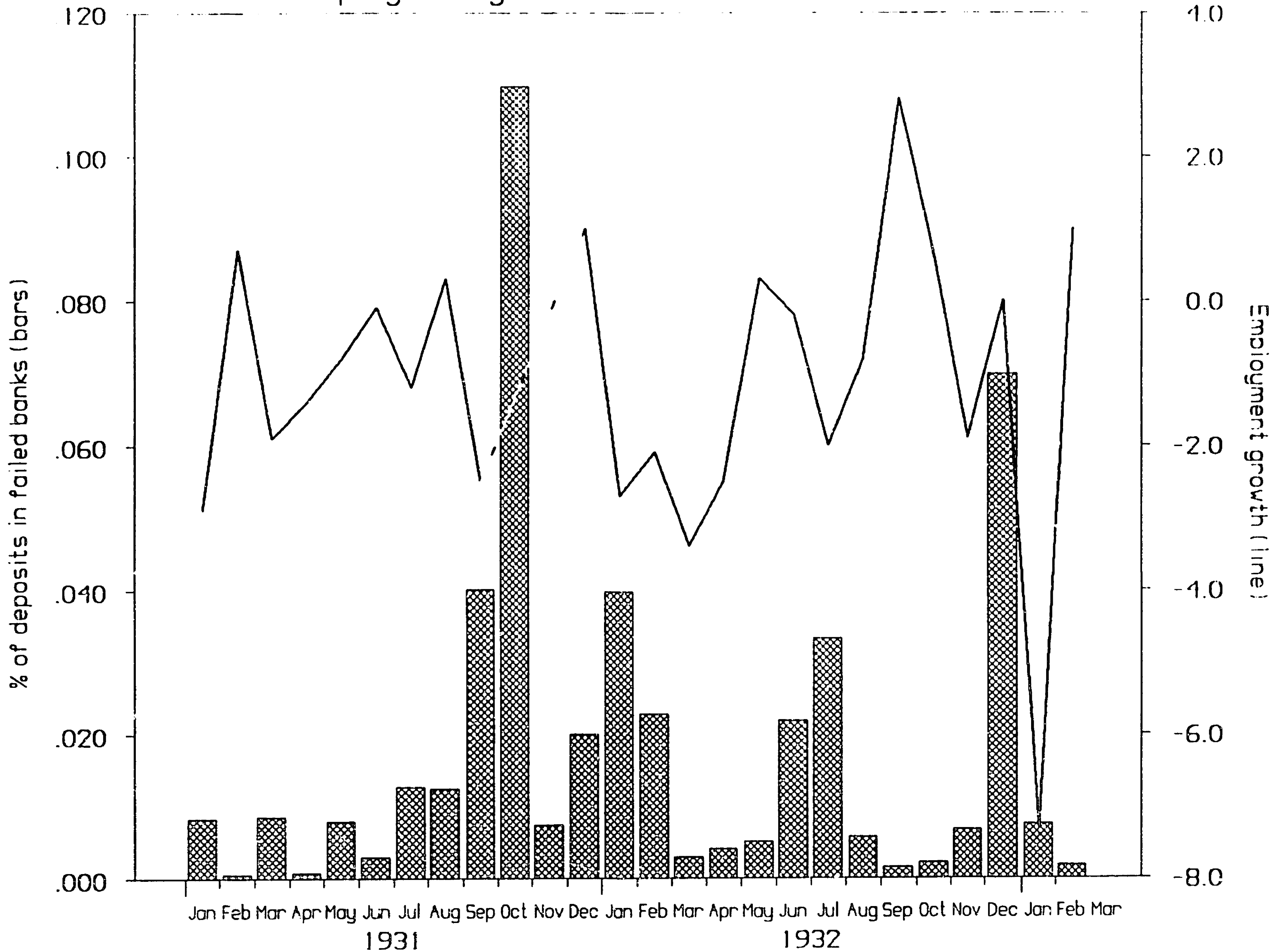

Industrial production and bank failures
Industrial production, seasonally adjusted (line)
% of deposits in failed banks (bars)
80
70
60
50
40
30
20
10
.0140
.0120
.0100
.0080
.0060
.0040
.0020
.0000
1921
1923
1925
1927
1929
1931
1933
1935
1937
1939

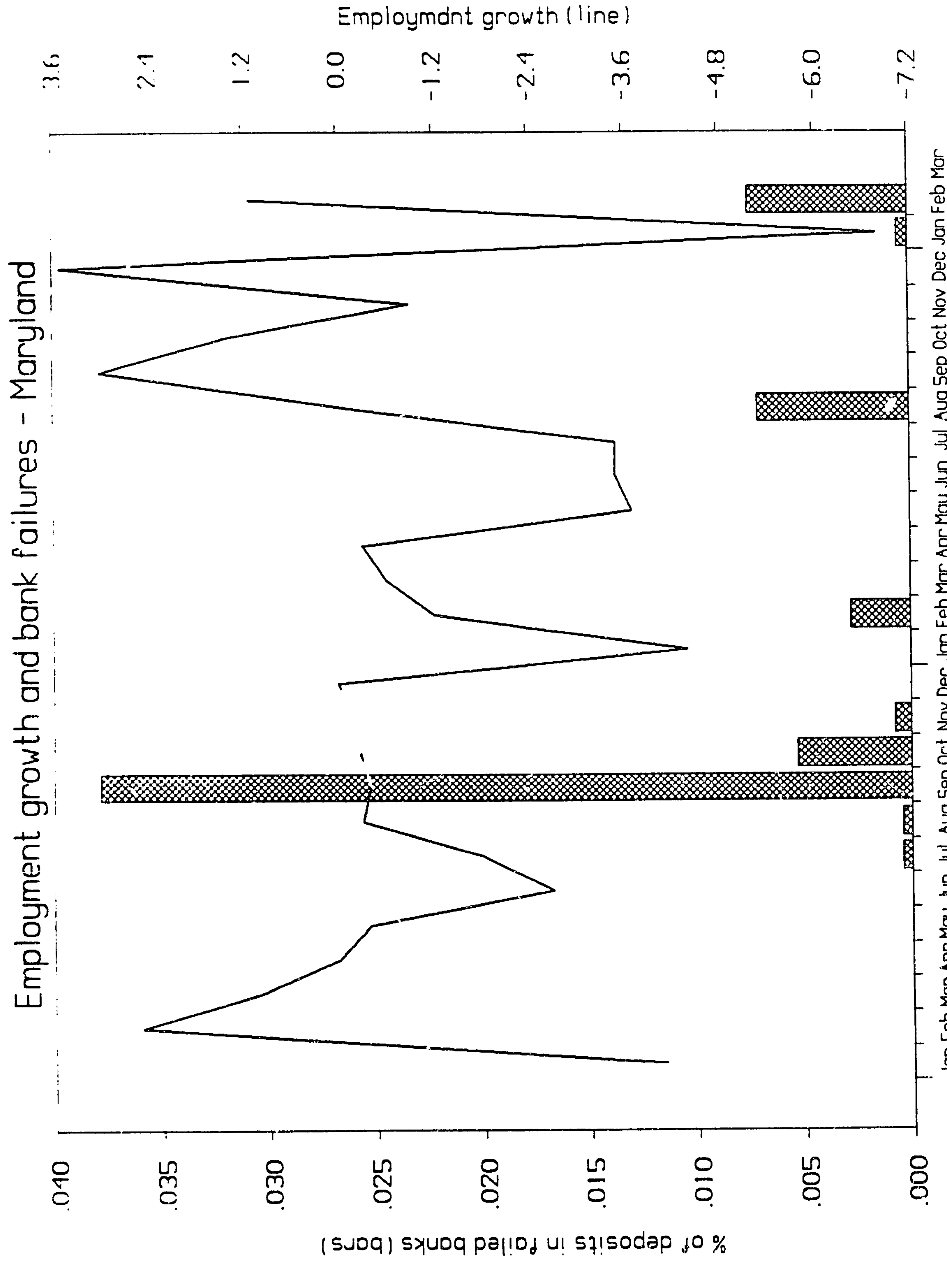

Employment growth and bank failures - California
.0125
.0100
.0075
.0050
.0025
.0000
% of deposits in failed banks (bars)
12.5
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
-7.5
Employment growth (line)
J F M A M J J A S O N D J F M A M J J A S O N D J F M
1931
1932



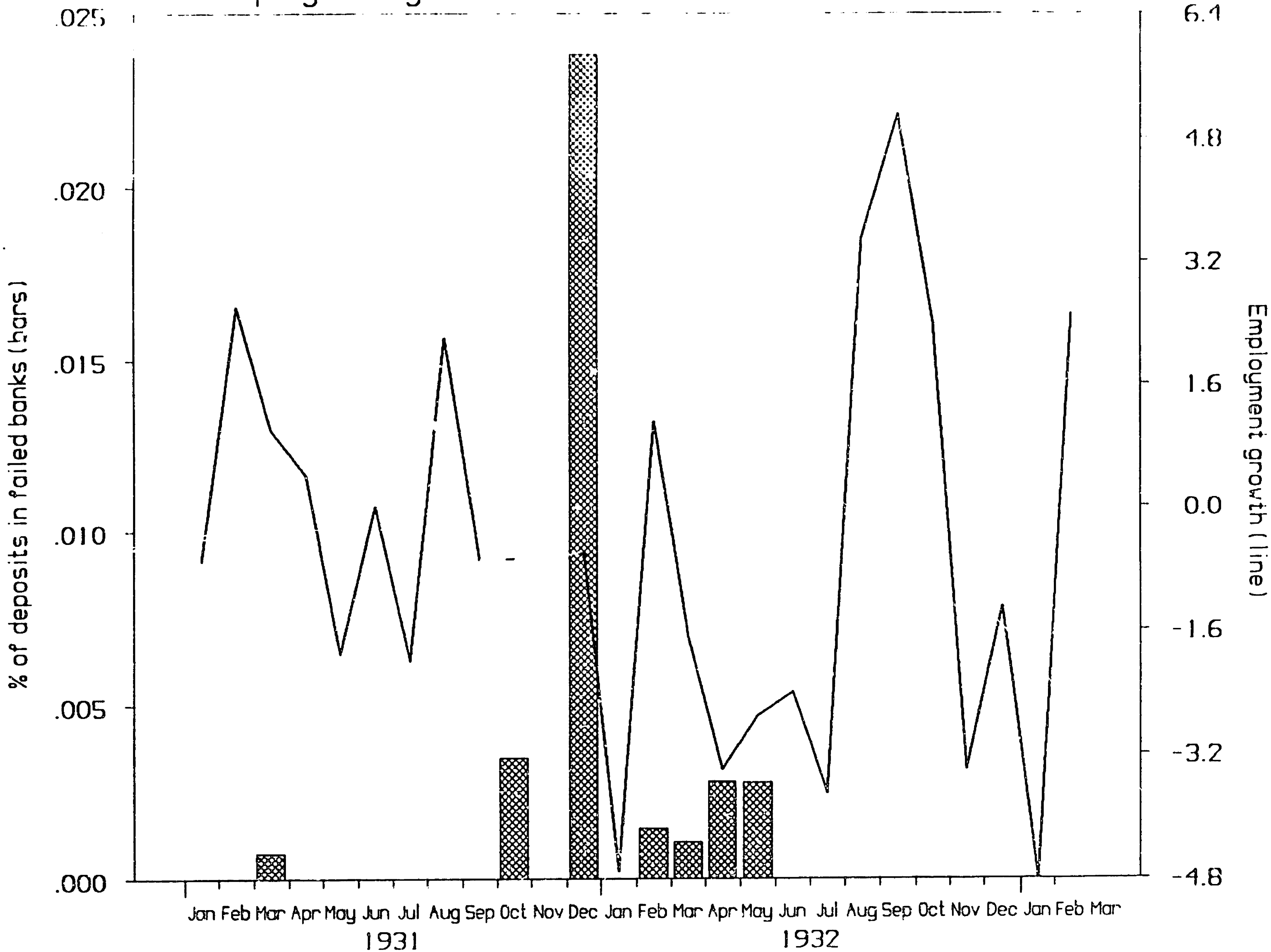
Employment growth and bank failures - Illinois
Employment growth (line)
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
.035
.030
.025
.020
.015
.010
.005
.000
% of deposits in failed banks (bars)
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar
1931
1932

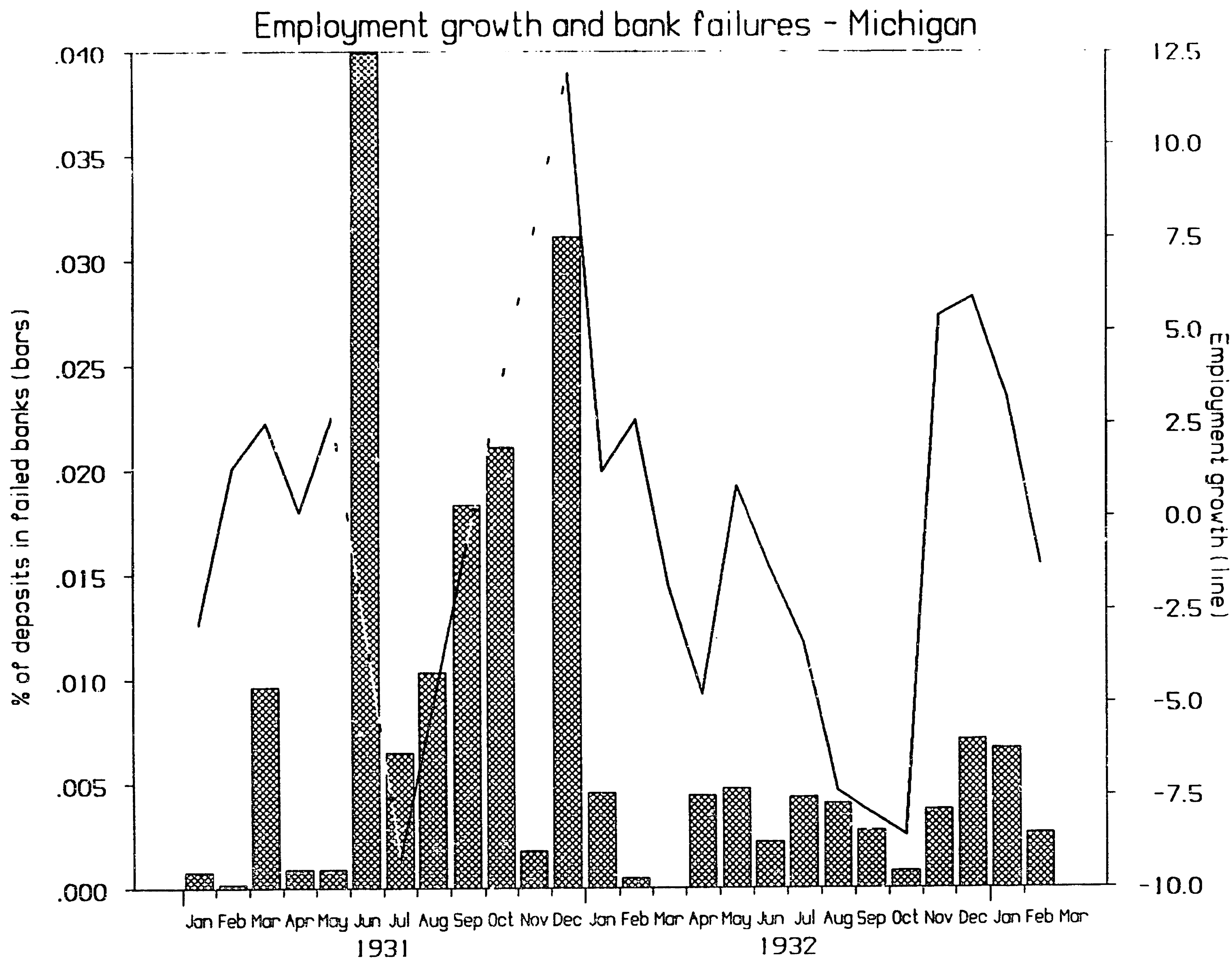
# Employment growth and bank failures - Iowa



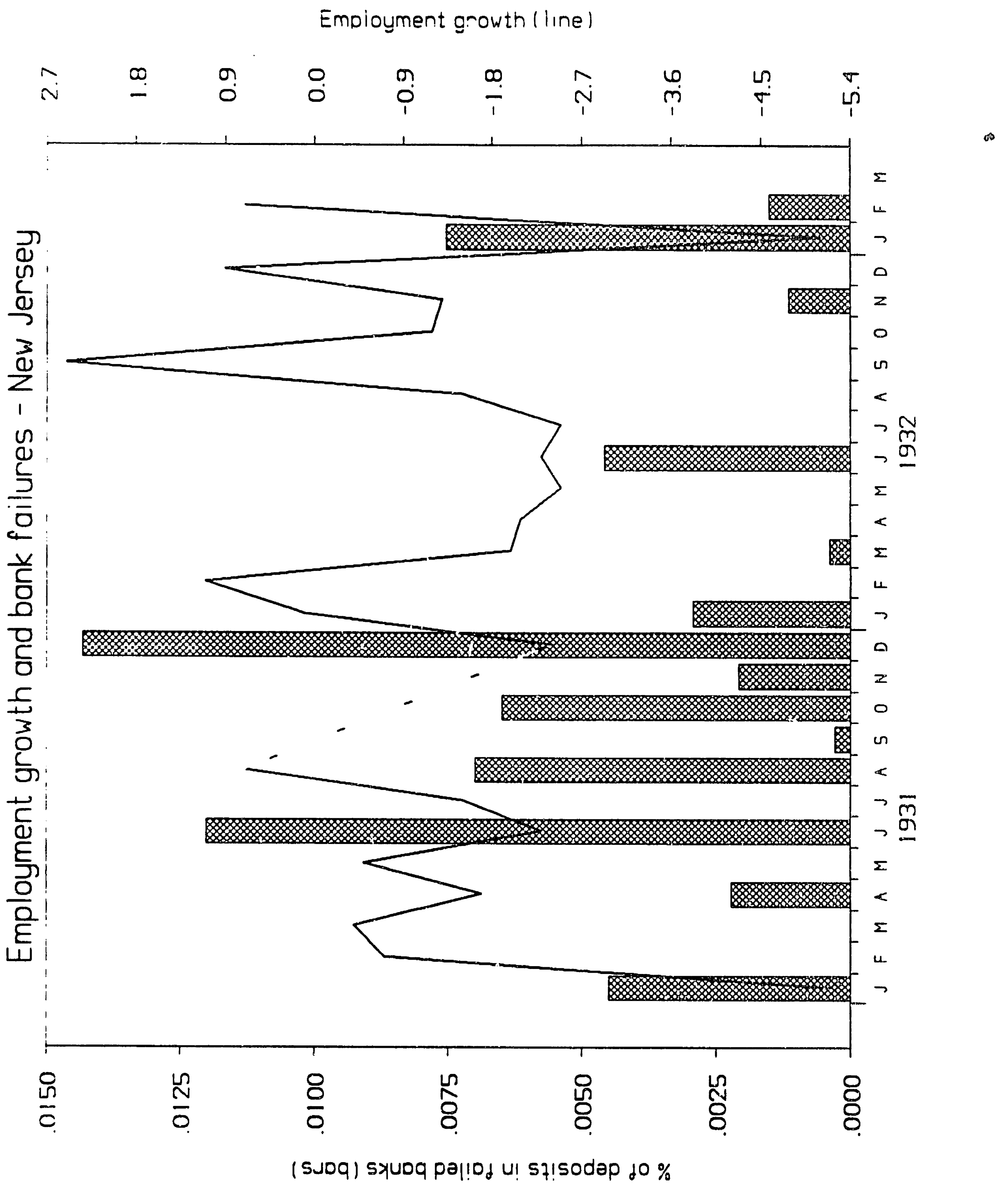

Employment growth and bank failures - Maryland
Employmdnt growth (line)
3.6
2.4
1.2
0.0
-1.2
-2.4
-3.6
-4.8
-6.0
-7.2
% of deposits in failed banks (bars)
.040
.035
.030
.025
.020
.015
.010
.005
.000
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar
1931
1932

Employment growth and bank failures – Massachusetts

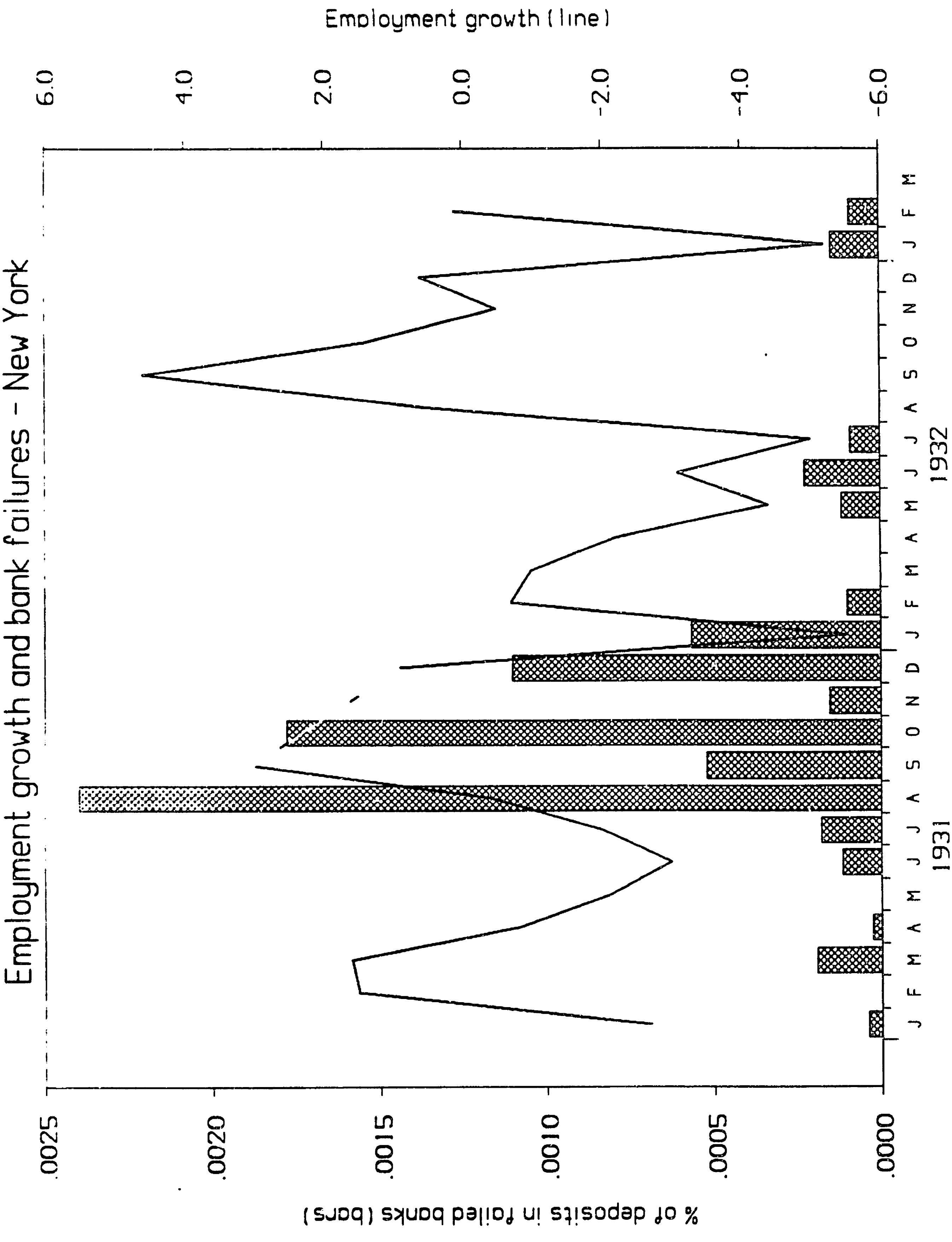


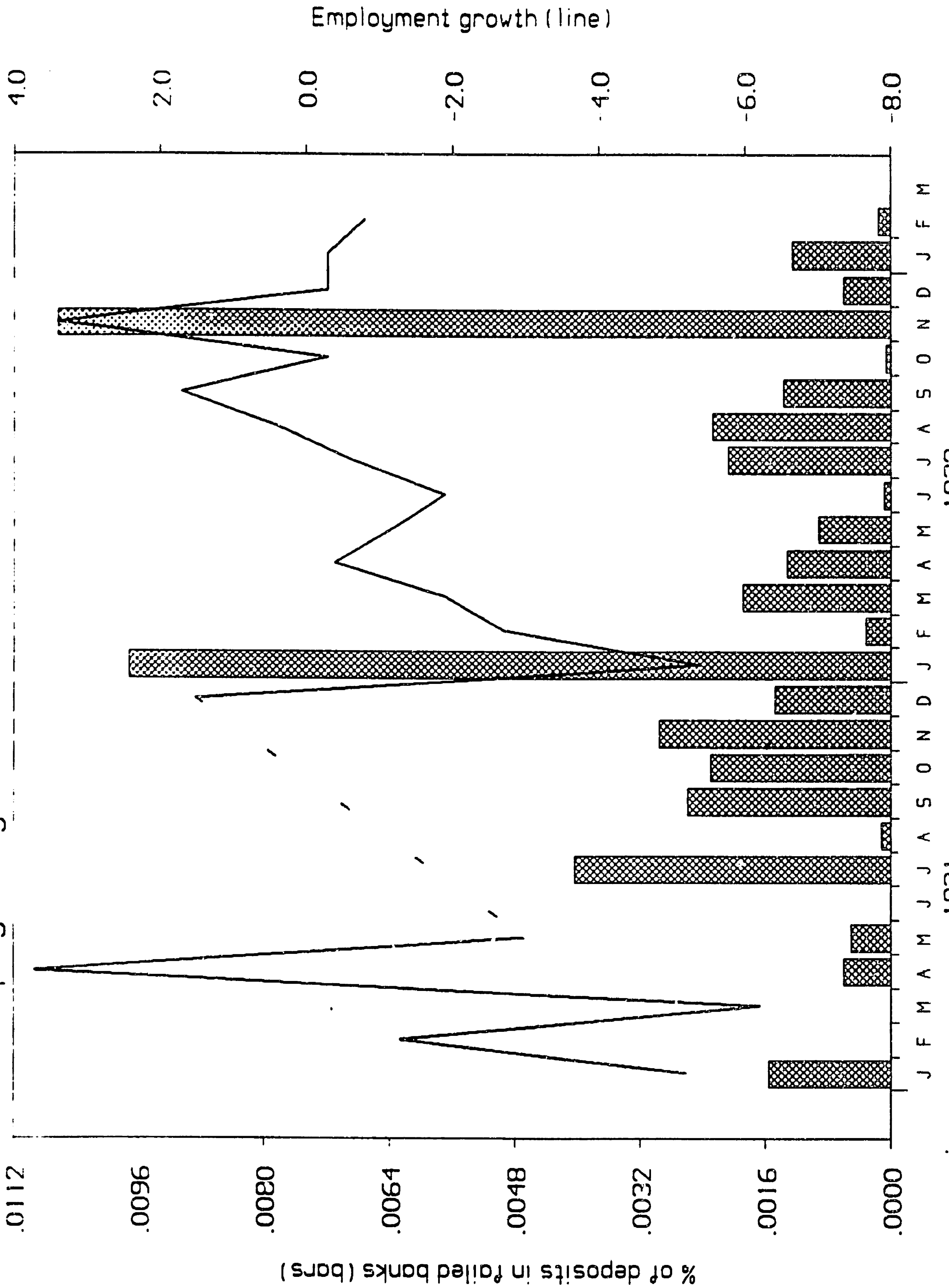


Employment growth and bank failures - Michigan
% of deposits in failed banks (bars)
.000
.005
.010
.015
.020
.025
.030
.035
.040
Employment growth (line)
12.5
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
-7.5
-10.0
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar
1931
1932

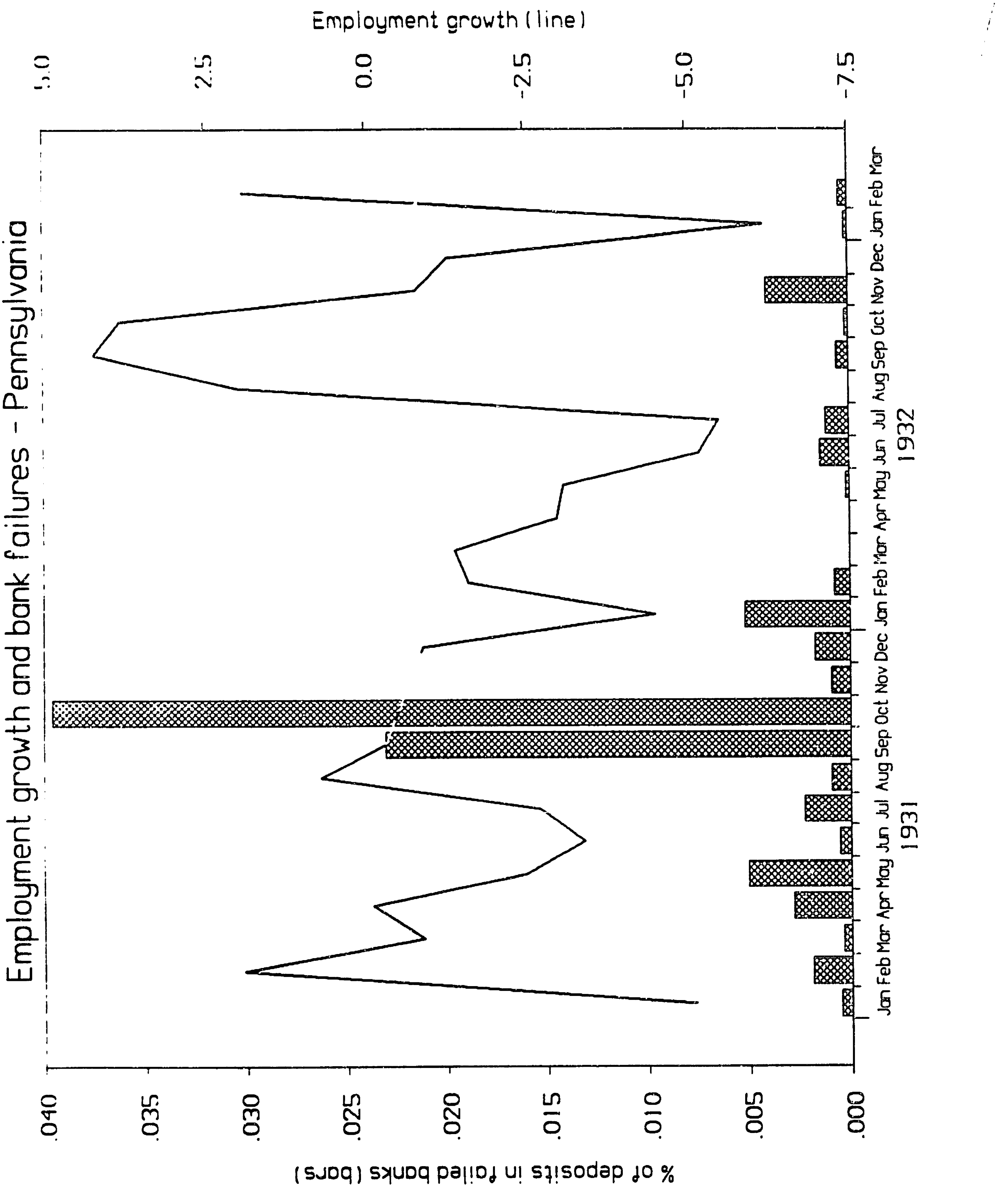


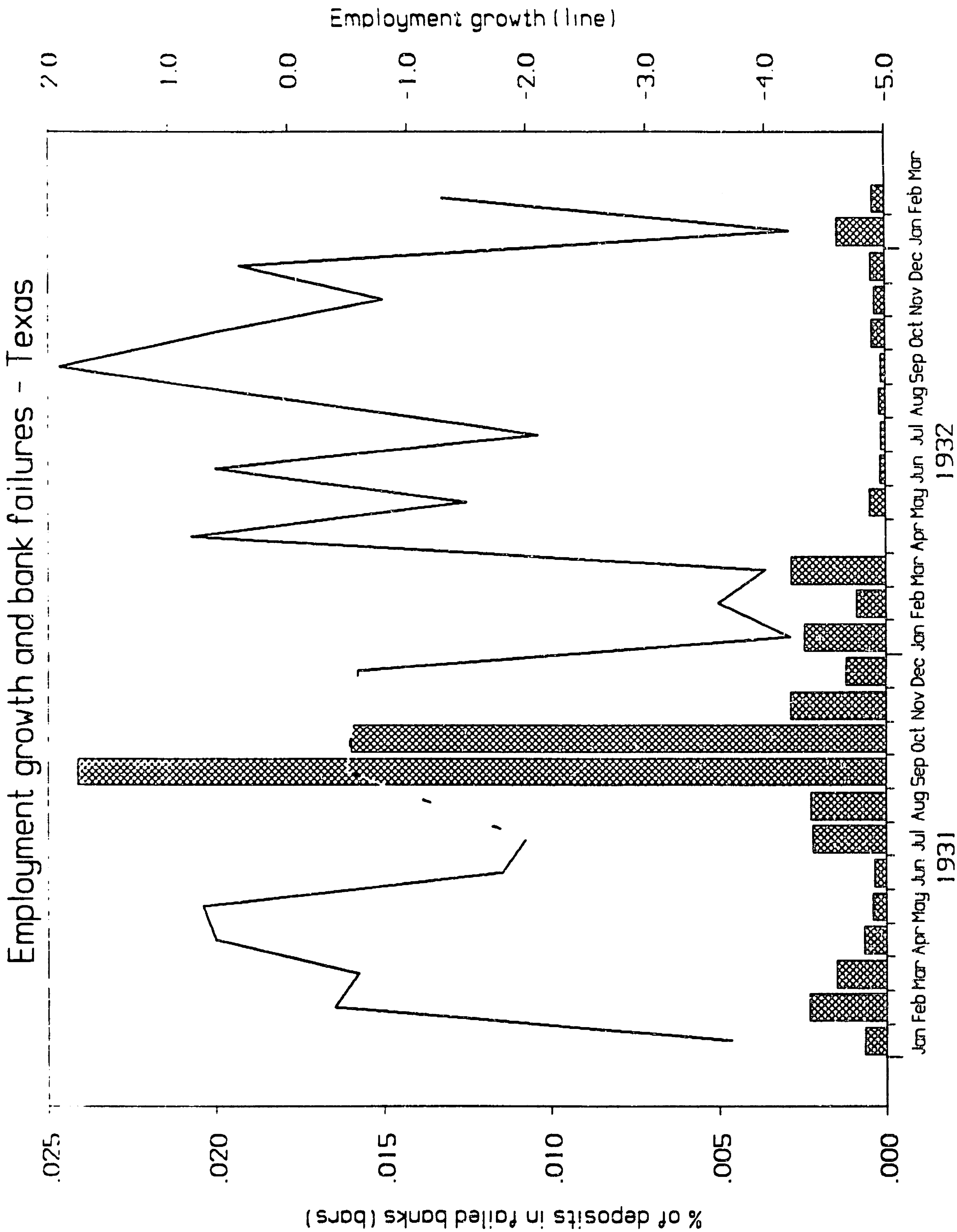
Employment growth and bank failures - New Jersey
Employment growth (line)
2.7
1.8
0.9
0.0
-0.9
-1.8
-2.7
-3.6
-4.5
-5.4
% of deposits in failed banks (bars)
.0150
.0125
.0100
.0075
.0050
.0025
.0000
J F M A M J J A S O N D J F M A M J J A S O N D J F M
1931
1932


Employment growth and bank failures - New York
Employment growth (line)
6.0
4.0
2.0
0.0
-2.0
-4.0
-6.0
% of deposits in failed banks (bars)
.0025
.0020
.0015
.0010
.0005
.0000
J F M A M J J A S O N D J F M A M J J A S O N D J F M
1931
1932

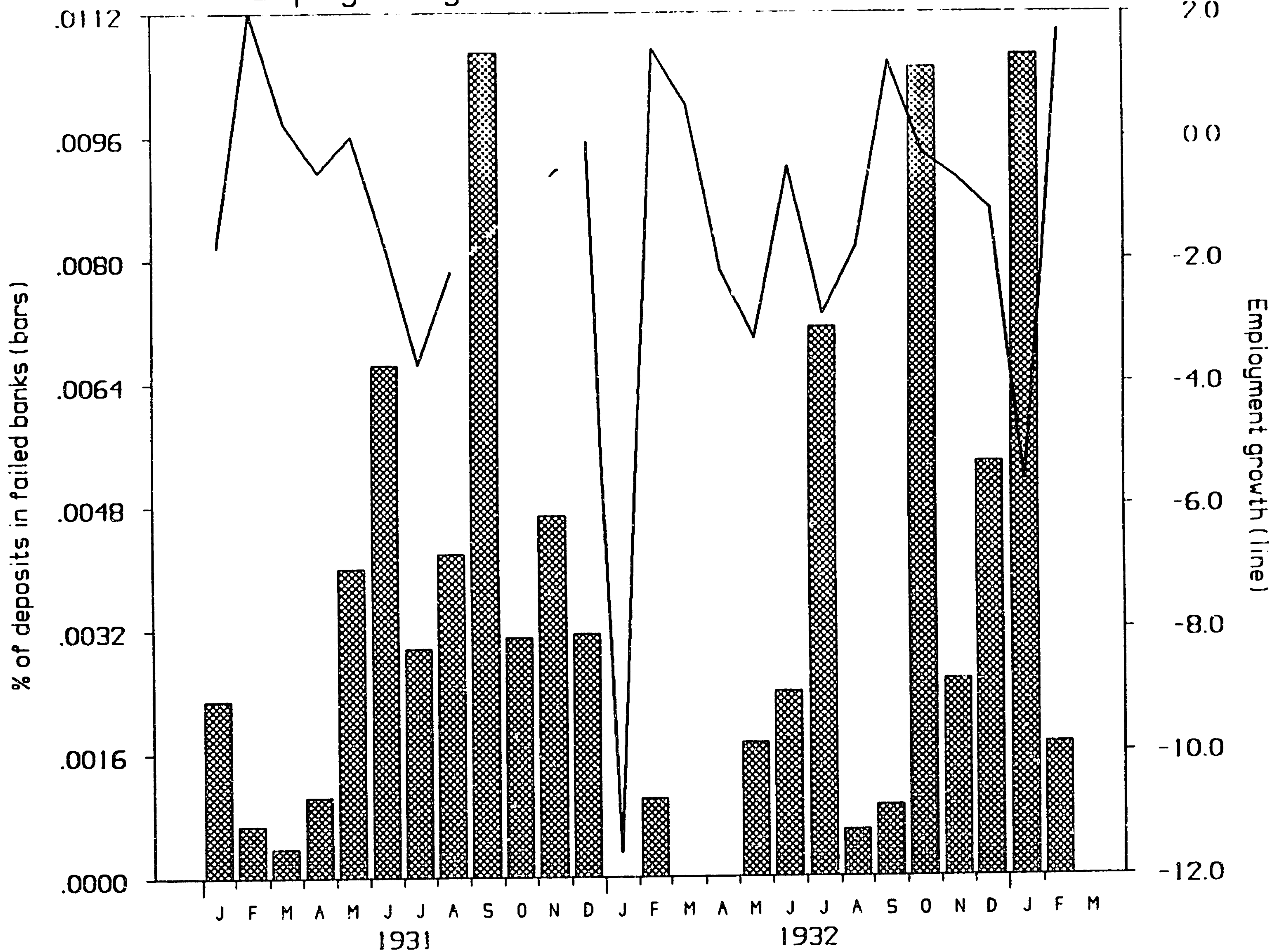



Employment growth and bank failures - Oklahoma


Employment growth and bank failures - Pennsylvania
Employment growth (line)
2.5
0.0
-2.5
-5.0
-7.5
% of deposits in failed banks (bars)
.040
.035
.030
.025
.020
.015
.010
.005
.000
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar
1931
1932


Employment growth and bank failures - Texas
Employment growth (line)
2.0
1.0
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
% of deposits in failed banks (bars)
.025
.020
.015
.010
.005
.000
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar
1931
1932

Employment growth and bank failures - Wisconsin



# References


*Annual Statistical Digest*, Board of Governors of the Federal Reserve System, various issues.

*Banking and Monetary Statistics*, Board of Governors of the Federal Reserve System, 1943.

*Banking and Monetary Statistics 1941–1970*, Board of Governors of the Federal Reserve System, 1976.

Bentolila, Samuel and Giuseppe Bertola, "Firing Costs and Labour Demand: How Bad is Eurosclerosis?" *Review of Economic Studies* 57 (1990), 381–402.

Bernanke, Ben S., "Nonmonetary Effects of the Financial Crisis in the Propagation of the Great Depression," *American Economic Review* 73:3 (June 1983), 257–275.

Bernanke, Ben and Alan Blinder, "The Federal Funds Rate and the Channels of Monetary Transmission," NBER Working Paper No. 3487, October 1990.

Bernanke, Ben and Mark Gertler, "Agency Costs, Net Worth, and Business Fluctuations," *American Economic Review* 79 (March 1989), 14–31.

Bertola, Giuseppe, *Adjustment Costs and Dynamic Factor Demands: Investment and Employment under Uncertainty*, Ph.D. Thesis, MIT, 1988.

Bertola, Giuseppe and Ricardo Caballero, "Kinked Adjustment Costs and Aggregate Dynamics," in O.J. Blanchard and S. Fischer, eds., *NBER Macroeconomics Annual 1990*, volume 5, MIT Press, 1990

Bisignano, Joseph, "Structures of Financial Intermediation, Corporate Finance and Central Banking," Bank for International Settlements mimeo, December 1990.

Blackwell, Norman R. and Anthony M. Santomero, "Bank Credit Rationing and the Customer Relation," *Journal of Monetary Economics* 9 (January 1982), 121–129.

Blanchard, Olivier Jean and Lawrence Katz, "Regional Evolutions," *Brookings Papers on Economic Activity*, 1992:1, 1–76.

Blanchard, Olivier and Mark Watson, "Are All Business Cycles Alike?" in R. Gordon, ed., *The American Business Cycle: Continuity and Change*, University of Chicago Press for NBER, 1986.

Blinder, Alan S., "Credit Rationing and Effective Supply Failures," *Economic Journal* 97 (June 1987), 327–352.

Calomiris, Charles W., "Is Deposit Insurance Necessary? A Historical Perspective," *Journal of Economic History* 50:2 (June 1990), 283–295.

Caplin, Andrew and John Leahy, "State-Dependent Pricing and the Dynamics of Money and Output," *Quarterly Journal of Economics* 106 (August 1991), 683–708.

Constantinides, George M. and Scott F. Richard, "Existence of Optimal Simple Policies for Discounted-Cost Inventory and Cash Management in Continuous Time," *Operations Research* 26:4 (July-August 1978), 620–636.

Diamond, Douglas W., "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies* 51 (1984), 393–414.

Díaz-Giménez, Javier, Edward C. Prescott, Terry Fitzgerald, and Fernando Alvarez, "Banking in Computable General Equilibrium Economies," Federal Reserve Bank of Minneapolis Research Department Staff Report 153, March 1992.

Dixit, Avinash, "Entry and Exit Decisions under Uncertainty," *Journal of Political Economy* 97 (June 1989), 620–638.

Dumas, Bernard, "Super Contact and Related Optimality Conditions," NBER Technical Working Paper No. 77, April 1989.

Eckstein, Otto and Allen Sinai, "The Mechanisms of the Business Cycle in the Postwar Era," in R. Gordon, ed., *The American Business Cycle: Continuity and Change*, University of Chicago Press for NBER, 1986.

Fama, Eugene F., "What's Different about Banks?" *Journal of Monetary Economics* 15 (1985), 29–39.

Farrell, Joseph and Carl Shapiro, "Optimal Contracts with Lock-In," *American Economic Review* 79 (March 1989), 51–68.

*Federal Reserve Bulletin*, various issues.

Friedman, Milton, "Comment on Tobin," *Quarterly Journal of Economics* 84 (May 1970), 318–327.

Friedman, Milton, and Anna Jacobsen Schwartz, *A Monetary History of the United States*, Princeton University Press, 1963.

Gertler, Mark, "Financial Capacity, Reliquification and Production in an Economy with Long-term Financial Arrangements," NBER Working Paper No. 2763, November 1988.

Godfrey, L.G., *Misspecification Tests in Econometrics*, Cambridge University Press, 1988.

Harrison, J. Michael, *Brownian Motion and Stochastic Flow Systems*, John Wiley and Sons, 1985.

Haubrich, Joseph G., "Financial Intermediation, Delegated Monitoring and Long-Term Relationships," *Journal of Banking and Finance* 13 (1989), 9–20.

Hausman, Jerry A., "Specification Tests in Econometrics," *Econometrica* 46 (November 1978), 1251–1271.

Hodgman, Donald R., *Commercial Bank Loan and Investment Policy*, Bureau of Economic and Business Research, University of Illinois, 1963.

*Industrial Production*, Board of Governors of the Federal Reserve System, 1962.

Jaffee, Dwight M., *Credit Rationing and the Commercial Loan Market*, John Wiley and Sons, 1971.

Kane, Edward J. and Burton G. Malkiel, "Bank Portfolio Allocation, Deposit Variability, and the Availability Doctrine," *Quarterly Journal of Economics* 79 (1965), 113–134.

Keeton, William, *Equilibrium Credit Rationing*, Garland Press, 1979.

Minsky, Hyman P., *Stabilizing an Unstable Economy*, Yale University Press, 1986.

*Monthly Labor Review*, various issues.

Morgan, Donald P., "Financial Contracts When Costs and Returns Are Private: Why a Line of Credit Beats a Loan," Federal Reserve Bank of Kansas City mimeo, January 1992.

Prescott, Edward, "Theory Ahead of Business Cycle Measurement," *Federal Reserve Bank of Minneapolis Quarterly Review*, Fall 1986, 9–22.

Romer, Christina D. and David H. Romer, "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz," in O. J. Blanchard and S. Fischer, eds., *NBER Macroeconomics Annual 1989*, volume 4, MIT Press, 1989.

Romer, Christina D. and David H. Romer, "New Evidence on the Monetary Policy Transmission Mechanism," *Brookings Papers on Economic Activity* 1990:1, 149–213.

Secrist, Horace, *National Bank Failures and Non-Failures*, Principia Press, 1938.

Sharpe, Steven A., "Asymmetric Information, Bank Lending, and Implicit Contracts: A Stylized Model of Customer Relationships," *Journal of Finance* 45 (September 1990), 1069–1087.

Sofianos, George, Paul Wachtel, and Arie Melnik, "Loan Commitments and Monetary Policy," *Journal of Banking and Finance* 14 (1990), 677–689.

Stiglitz, Joseph E. and Andrew Weiss, "Credit Rationing in Markets with Imperfect Information," *American Economic Review* 71 (June 1981), 393–410.

Temin, Peter, *Did Monetary Forces Cause the Great Depression?*, W.W. Norton & Company, 1976.

Temin, Peter, *Lessons from the Great Depression*, MIT Press, 1989.

Tirole, Jean, *The Theory of Industrial Organization*, MIT Press, 1988.

Tobin, James, "Money and Income: *Post Hoc Ergo Propter Hoc?*" *Quarterly Journal of Economics* 84 (May 1970), 301–317.

Townsend, Robert M., "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory* 21 (1979), 265–293.

Wallis, John Joseph, "Employment in the Great Depression: New Data and Hypotheses," *Explorations in Economic History* 26 (1989), 45–72.

Williamson, Stephen D., "Financial Intermediation, Business Failures, and Real Business Cycles," *Journal of Political Economy* 95 (December 1987), 1196–1216.

Wojnilower, Albert, "The Central Role of Credit Crunches in Recent Financial History," *Brookings Papers on Economic Activity*, 1980:2, 277–326.

Wood, J.H., *Commercial Bank Loan and Investment Behavior*, John Wiley and Sons, 1975.

# Biographical Note

Michael Scott Gibson was born May 12, 1966 in Santa Monica, California. He attended Stanford University from September 1984 to June 1988 and graduated with an A.B. degree in Economics, with distinction, with Honors in Humanities. He pursued a Ph.D. in the Department of Economics at the Massachusetts Institute of Technology from September 1988 to June 1992. From June through August 1990 he consulted at the Bank of Portugal in Lisbon. In July 1992 he joined the staff of the Board of Governors of the Federal Reserve System as an Economist in the Division of International Finance. In February 1993 he will be awarded his Ph.D. degree. In June 1993 he will marry Colette M. Hanson.