The Past and Future of Commercial Banking Viewed Through an Incomplete Contract Lens

Author(s): Raghuram G. Rajan


Published by: Ohio State University Press


Accessed: 30-11-2015 16:38 UTC
The Past and Future of Commercial Banking
Viewed through an Incomplete Contract Lens

Commercial banks emerged at a time when contracts were very incomplete and property rights insecure. They typically offered demand deposits, made loans on demand, and were regulated. Each of these aspects of the institutional structure were essential in helping the bank provide the twin functions of liquidity and safety. I argue that recent theories of banking, which I collectively refer to as “Incomplete Contract” theories of banking, explain well the origins of banking. I also claim that they can explain recent changes in banking: as the informational, legal, and property rights environment has improved, there appear to be fewer synergies between various aspects of the traditional institutional structure of the bank. In developed countries, it is now time to think whether there is anything special about the institutional form of the bank, or whether all that is special is that it is regulated.

According to many observers, the commercial bank—the institution that accepts deposits payable on demand and originates loans—has outlived its usefulness and is in a state of terminal decline. Commercial banks’ share of total assets in financial institutions in the United States has fallen dramatically, from over 70 percent around the turn of the century to just around 30 percent today (see Boyd and Gertler 1994; James and Houston 1996). Bank share of corporate debt in the United States has declined from 19.6 percent in 1979 to 14.5 percent in 1994 (Berger, Kashyap, and Scalise 1995). Competition on both sides of the bank’s balance sheet has increased. On the asset side, the growth of the commercial paper and junk bond markets has given large firms an alternative to borrowing from the bank. On the liability side, new technologies and deregulation have given customers choice. Instead of being forced to deposit at the local bank branch, or make payments through a bank checking account, a mutual fund which offers much the same services is just a dial tone away. Given these changes, it is, therefore, legitimate to ask: are commercial banks dying?

Theoretically, this is tantamount to asking two questions. First, are the functions that the bank has historically performed obsolete? Second, is the institutional form

This paper is, in part, a distillation of conversations between the author and Anil Kashyap, Randall Kroszner, Mitchell Petersen, Stewart Myers, Jeremy Stein, and especially Douglas Diamond and Luigi Zingales. Some of the ideas in this paper originated during a stay at the Stockholm School of Economics which the author thanks for its hospitality. He is also grateful to the NSF for its support (grant SBR 9423645).

1. Also see Gorton and Rosen (1995).

RAGHURAM G. RAJAN is the Joseph L. Gidwitz Professor of Finance at the University of Chicago’s Graduate School of Business.
Journal of Money, Credit, and Banking, Vol. 30, No. 3 (August 1998, Part 2)
that carried out these functions no longer useful? While traditional theories have given us a sense of the functions banks perform, they are less specific about why these functions are provided by the institutional form that most of us refer to as a commercial bank. In this sense, traditional theories are not very useful in helping us understand whether banks are an anachronism.

Before I go further, let me be more specific about the institutional form under investigation. The U.S. Banking Act of 1971 defines the commercial bank as an institution that offers demand deposits and originates loans. Therefore, a money market mutual fund is not a bank (it does not originate loans) and a finance company is also not a bank (it does not offer demandable deposits). In addition to the products offered, an important aspect of the institutional form has been that banks are typically regulated, have government guarantees, and have access to a public lender of last resort.

My first task is to analyze the functions the bank has historically performed through this institutional form. I will specifically ask why the products of demand deposits and loan origination are combined under one roof, and why this roof has been regulated. Given space constraints, the historical facts and sources I draw on are undoubtedly shaped by my theoretical biases. Given that I am not a historian and few historians pay attention to banking theory, I justify this on grounds that it seems the only way to move ahead.

I argue that banks arose historically to provide customers liquidity and a safe investment haven. This is not a particularly novel position, though the details of my observations differ in important ways from the received wisdom in the literature. What I want to emphasize is that given the poor legal environment, and the lack of institutional checks and balances on power, contracts were very incomplete and property rights insecure. The institutional form of the bank, I argue, arose to improve transaction possibilities over what was contractible through the market place. I suggest that regulation sometimes helped in this, and should be thought of as an integral part of the institutional structure of banks.

Many of the arguments that I will put forward reflect recent work (for example, Boot, Greenbaum, and Thakor 1993; Diamond and Rajan 1997; Kiyotaki and Moore 1996; and Myers and Rajan 1997) that are representative of what could be called the “incomplete-contract” approach to banking. These theories build upon previous work but alter the focus. Prior literature focuses more on the functions of the bank and less on why the institution works differently from contracts available in the market. An example may make the point clear. The bank in probably the most influential recent paper on banking, by Diamond and Dybvig (1983), is essentially an ex ante insurance contract assuring the signers of liquidity. While the Diamond and Dybvig model implies an institutional structure, it is not the focus of that paper. This has been a source of controversy because Jacklin (1987) shows that other contracts can replicate the outcomes in Diamond and Dybvig but these contracts look much less like a bank [though see Diamond (1997)]. While indeed the bank may perform the function of liquidity insurance, there must be something in the institutional structure that enables it to per-

2. See Bhattacharya and Thakor (1993) for an excellent survey of traditional theories of banking.
form this function better than direct contracting in the market place. Diamond and Rajan (1997) show that this “something” might lie in the details of the relationship between the banker, depositor, and borrower.

By focusing on the noncontractual mechanisms that institutions bring to bear, the “incomplete-contract” approach may shed more light on why the banking institution has not been easily replicated by the market in the past, and why its form is changing.\(^3\) In the incomplete-contract approach, the bank (as any other institution such as the firm) is more than simply a “nexus of contracts.” It exists precisely because the incompleteness of contracts allows it to bring a variety of noncontractual mechanisms into play. For instance, the distinguishing feature of the bank may not simply be the contracts it writes with depositors and borrowers. Instead, it may be its investment in reputation (Book, Greenbaum, and Thakor 1993), its investment in relationships with clients (Diamond and Rajan 1997), or the investment of its employees to one another and to the business. The point is that these institutional abilities cannot be replicated instantaneously in the market and, therefore, are what give the institution content. In situations of contractual incompleteness, they interact with the contracts that can be written to produce outcomes—in terms of liquidity and safety—that can be better than what the market can produce.

This puts the rationale for banking institutions squarely on contractual incompleteness. A secular change in institutional form can then be explained by a secular change in the degree of contract incompleteness. Specifically, I argue that the information and technology revolution has been accompanied by dramatic changes in the legal environment, all of which have enhanced contractual possibilities. The institutional form that traditionally provided the functions of liquidity and safety may not dominate other forms in this environment. This is not to say that the functions of liquidity provision and safety are no longer necessary, only that the form that optimally provides them can be very different from the one that provided them in the past. I also present some thoughts on what regulation is necessary in this improved contractual environment.

Finally, a methodological issue: Some argue that the incomplete-contract approach is less fundamental than an approach that starts from tastes and preferences. This criticism is hard to understand. The incomplete-contract approach does analyze outcomes based on agents’ tastes and preferences. One might argue that it makes assumptions about the capacity to contract and treats them as exogenous and on the same plane as assumptions about tastes and preferences. While this is true, it is no different from what the traditional literature does. The traditional literature implicitly makes assumptions about the capacity to contract—effectively assuming it to be limitless. The incomplete-contract literature makes explicit assumptions about the capacity to contract, essentially placing limits on it. While one could argue that some of these assumptions are ad hoc, they are no less ad hoc than assumptions about limitless contractual possibilities. Ultimately, one has to endogenize all contracts and their enforcement mechanisms. Since this is too complicated at present, one is reduced to us-

\(^3\) The Incomplete Contract approach derives from seminal work by Grossman and Hart (1986) and Williamson (1975, 1985) which has been focused on the theory of the firm. The arguments on the distinction between institutions and markets are explored in greater detail in Rajan and Zingales (1997).
ing plausibility as a test for what makes sense. Clearly, an assumption of incomplete contracts in medieval Europe is more sensible than assuming full contractual enforce-

The rest of this paper is as follows. In section 1, I provide a historical overview of the functions banks performed, and in section 2, I discuss some rationales for why they were regulated. In section 3, I examine important changes in the contractual and property rights environment which, I argue in section 4, have reduced the synergies between various aspects of the traditional institutional form of banks. In section 5, I ask what this implies for what financial institutions should be regulated, and how. Section 6 concludes.

1. A SELECTIVE DIGRESSION INTO HISTORY

Why Deposits and Overdrafts?

The consensus among economic historians [this subsection draws from Kashyap, Rajan, and Stein (1998) and Meyers and Rajan (1998)] is that deposit banking in continental Europe evolved from the activity of money changing. The early Middle Ages saw an increasing use of coins rather than barter in trade. There was, however, a prob-

lem with the available coins. Coinage was imperfect, so coins could contain very dif-

ferent quantities of metal even when newly produced by the same mint. The money

changer specialized in coins so when a trader brought in coins, the money changer

could open an account for the trader rather than giving him new coins. If the trader

wanted to pay a supplier money, the money changer would simply make an account-

ing entry, debiting the trader’s account. If the supplier had an account with the money

changer, the money changer would credit the account, thus reducing the entire pay-

ment transaction to pen strokes.4

The money changer had to maintain a reserve of coins so as to make net payments to other bankers and to meet withdrawals by depositors. But not all the cash that was initially deposited had to be maintained as reserve since only a fraction of depositors would need their money at any time. Fractional reserve theories of banking suggest that banks channeled idle cash into loans to entrepreneurs. However, one can go fur-

ther and say that the nature of the loans made, as well as the identity of the recipients, was determined by the deposit business.

Banks typically did not make long-term loans. Instead, the early private banks allowed depositors to borrow by overdrawing their account. These over drafts were thus loans obtained virtually on demand by depositors and largely went to finance trade. From the perspective of the money changer, the overdraft facility (or its modern equivalent, the line of credit) was essentially the same as a deposit. Both products re-

quired the money changer to come up with cash on demand, that is, they were prod-

ucts through which the money changer provided liquidity. While with the overdraft

4. It is straightforward to see why deposits were denominated in nominal terms and were not a function of the bank’s asset value. Clearly, the primitive accounting system and the variability of coinage meant that a market value for the bank was not only very imprecise, but also very costly to establish.
facility, the money changer was not legally required to make the loan (he could refuse to allow the overdraft), in practice the difference was probably moot; once customers came to rely on the money changer’s overdraft facility, it would hurt his reputation almost as much if he refused to allow it without good reason, as it would if he refused to pay out on deposits.

Did the simultaneous offering of demand deposits and overdrafts then not make the bank more fragile? The answer is in the affirmative only if deposit withdrawals and borrowing on overdrafts are strongly positively correlated. Otherwise, Kashyap, Rajan, and Stein (1998) argue that by diversifying across the liquidity demands of depositors and borrowers, the money changer could amortize the fixed costs of obtaining the liquidity needed to service these demands (the fixed costs being, for example, the idle reserves, security arrangements to protect them, the arrangements for securing liquidity from other bankers, etc).

Why Loans to Depositors?

Why did the money changer offer these implicit loan commitments largely to depositors? Public information about any kind of merchant was scarce at that time. Furthermore, merchants had little collateral to offer. Since no lender wants to offer money without collateral or without adequate information on the creditworthiness of the borrower, the lack of public information made it very difficult for merchants to get credit from the wealthy. Money changers could play a critical role here since through a merchant’s deposit account, they obtained a great deal of information about the merchant’s business and creditworthiness [see Nakamura (1988) for a theoretical analysis of this point]. Lamoreaux (1994, p. 101) quotes an early New England banker thus:

the best paper to accept is that offered by firms or individuals who are in the habit of carrying balances with their bank from whom the accommodation should be obtained.

There appears to me no better means to determine the amount of risk a bank incurs than by regulating its loans according to the average balance carried.

Depositors were therefore natural candidates for loans. Also, the promise of loans when in need would make the depositor more eager to establish a good credit history with the money changer. This would assure the money changer of deposit inflows whenever the depositor was flush with cash. Thus a valuable quid pro quo relationship could be established with depositors through the overdraft facility.

Why Not Hundred Percent Reserve Banking?

Early banks were not models of sound banking practice. They failed often (see Riu 1979). Why did the money changer not let the money lie “idle” in his strong box, and charge the depositor an additional fee for providing riskless payment services? In other words, why did the equivalent of today’s money market mutual fund not start up?

5. Also see Diamond and Dybvig (1983) and Holmstrom and Tirole (1996, 1997) for a theoretical exposition of banks as institutions that diversify across liquidity demands. These papers, however, look at diversification only on one side of the balance sheet.
Why was 100 percent reserve banking, the solution to financial fragility so beloved of economists (for example, see Simons 1948), not a viable institutional form? One answer may simply be that it was too costly to let reserves lie idle. Goodhart (1995, p. 7) reports that the forerunner of the Swedish Riksbank was organized on the basis of two supposedly separate departments, the loan department financing loans on the basis of longer-term deposits and capital, and the issue department supplying credit notes on the receipt of gold and specie. But even when the bank had been taken over by Parliament, a secret instruction authorized the advance by the issue department to the lending department of the funds at its disposal.

Since separation within a financial institution, even one run by Parliament, was not credible at that time, perhaps only institutions that completely refrained from lending were possible. But then this meant that depositors gave up their access to demandable credit. Given that credit and information about creditworthiness was scarce, depositors probably did not want to bank with someone who did not offer loans.

There is some evidence in support of this. After repeated failures of private banks in Barcelona in the late fourteenth century, the municipality set up a public bank. The first such bank was the Taula de Canvi (Bank of Deposit) in Barcelona which began to function in 1041 [see Usher (1943) and Riu (1979) for a detailed description]. The Taula offered deposit services to the citizens. It also received surplus taxes and lent money at moderate rates to the city. The Taula had two separate accounting systems. The deposits of individuals wary of private bankers were covered by cash on hand, while the deposits of businessmen and professionals were balanced by loans to the state and official organizations. The Taula was different from private banks in that it was guaranteed by the municipality and, importantly, it was not supposed to make loans to the public. Thus the Taula represented a move towards the ideal of secure 100 percent reserve banking that is so often advocated (then and now) as a solution to the instability of private banks. Moreover, it concentrated on payment services. If overdraft lending was separable from deposit banking, the Taula should have been a resounding success with depositors. It was not! As Usher (1943, p. 273) puts it,

When the project of the Bank [Taula] was first discussed, it was presumed that private persons would welcome the opportunity to deposit their funds in an institution whose liabilities were guaranteed by the city as well as by sureties usually given by bankers. The long record of the struggle between the Bank of Deposit and the private banks shows clearly that the Bank of Deposit could not compete for private business unless some legislative pressure was exerted and special privileges given the Bank of Deposit. The restrictions placed upon loans to private persons would furnish an adequate explanation of the difficulty to the modern student, but the city authorities do not seem to have been fully conscious of the importance of the restriction on the lending power of the Bank of Deposit. The removal of restrictions on loans to private persons seems never to have been discussed, though . . . some qualification [was] introduced in practice.

Thus Usher places the shortcomings of the municipal bank squarely on its reluctance to offer overdrafts. He hints at but does not explain why overdrafts were so crucial to the bank’s attracting depositors. Certainly, part of the explanation must be that depositors thought the availability of credit, as they built credit histories through their
deposit account, important. Given the restriction on lending by the public bank, the information from credit histories could only be utilized by the private bankers, hence their attraction.

Yet another reason why banks did not hold 100 percent reserves was that it just was not possible given the legal and property rights environment of that time; a bank may have found it very hard to inspire confidence when it was fully invested in cash [see Myers and Rajan (1998) for a theoretical discussion]. The difficulty in safely holding liquid cash may be hard to see from today’s vantage point when we have large money market mutual funds safely keeping billions of dollars of money-like assets in trust for individuals. What many of us do not appreciate are the enormous improvements in accounting, valuation, and information technology, as well as the development of institutions such as third-party custodians of securities that have made such funds possible. What is also not immediately apparent is the extensive regulation and scrutiny, as well as the respect for private property rights, that allows these institutions to inspire confidence.

Seven hundred years ago, it would have been impossible for depositors to have confidence in a money changer of significant size who kept all the deposited cash as specie on his premises. Unlike the net asset values that are calculated and published every day for modern funds, it would be very hard for depositors to verify specie holdings. Apart from the time and effort in counting the coins physically (computers and script-less depositories have clearly helped here) their inability to understand the value of the coins is what led them to entrust the money changer with the coins in the first place. Clearly, even though the money changer facilitated payments, depositors were concerned about the safety of their money, especially given the opacity of the money changer’s strong box.

Were depositors right to be wary? Detailed studies of the money changers in Catalonia in Spain from the thirteenth to fifteenth century suggest they were (Riu 1979, for example). The primary cause of bank failure seems to have been fraud. The money changers were repeatedly accused of making fraudulent transfers to fictitious persons and then declaring an inability to repay depositors.

Governments could also not be trusted to keep their hands off bank reserves. A number of bankers in Barcelona were executed by the city in 1345 even though they failed as a result of forced loans to the city itself (Riu 1979, p. 143). Executing the lender was a convenient way of avoiding repayment and improving city finances, a device many municipalities today probably regret having dispensed with. Sovereigns often demanded loans in exchange for continuing the privileges bankers enjoyed, and it was hard for a banker to refuse them especially when he was not already fully extended. Many banks including the famed Bardis and the Fuggers failed on the back of loans to sovereigns. Given the dangers of these loans it is natural that the money changers would attempt to find ways of refusing them. There is some evidence that the Medicis attempted to prohibit loans to sovereigns in the by-laws of their partnership agreements (De Roover 1968, p. 86). But these contractual covenants obviously did not provide sufficient protection against princely demands because the Medicis were sucked into making loans to the Duke of Milan which led to their eventual failure.
Money changers, therefore had to find a better way of assuring depositors of the safety of their money than locking it up in a strong box or simply stating in covenants that they would be prudent about the uses the money was put to.

Since the government could not be trusted to respect private property (especially in times of need such as war), and reputable third-party custodians did not exist, the money changer had to find someone else to be custodian of the excess reserves. This is where the overdraft may have been an appropriate institutional arrangement in a second-best world where property rights were not respected. Who better to target as custodians of the reserves than the depositors themselves? Since the money changer would keep only enough specie to meet the random needs of his borrowers and depositors, he could not misappropriate significant amounts of cash (relative to the profits generated by his business) without being detected. Furthermore, even if he chose to cut and run, he could only vanish with the fractional reserve since the rest of the assets would be locked up in overdrafts. Finally, with much of the money being widely distributed and with the remaining reserves being necessary to service deposits, reserves were much less attractive to local powers. It was harder for the sovereign to demand loans on grounds that the reserves were not immediately needed. Furthermore, a forced levy would likely force the bank into default and risk angering the typically rich and influential depositors.

I do not mean to suggest that money, once lent out, stays with the borrower and is not redeposited in the bank though the argument is structured that way. The essential point is that by fully utilizing reserves through overdrafts, the money changer could commit to protect depositors better against misappropriation. Moreover, he created a larger number of stakeholders who would help him in any stand against the sovereign (see Kindleberger 1993, p. 47). So the money changer aggregated both the financial and political power of the depositors, and focussed these toward their common interest.

There is some evidence for the idea that the fear of sovereigns was an impetus to the activities of banks. The natural precursor of the deposit bank, the money changer, did not exist in England since the Royal Mint monopolized the business of coin exchange (Lane and Mueller 1985, p. 65). Even though London goldsmiths accepted money in trust before the Civil War, this was not a general practice. Instead, the Mint in the Tower was typically used as repository for the excess specie and plate merchants had. But in 1640, Charles I confiscated 200,000 pounds in coin and bullion that had been deposited in the Mint by the London merchants. The act destroyed the Mint’s reputation as a safe depository. After strong protests going right up to the Privy Council, the King agreed to return the amount taken, on condition that the depositors granted him a loan of 40,000 pounds.

6. The essential idea is that the money changer expands the volume of activity through overdrafts so that the fractional reserves are just enough to sustain the business. At this point, it is hard to reduce the reserve without detection. Also, any such reduction will be extremely costly in terms of lost business. So the money changer has little incentive to misappropriate reserves. Finally, so long as the reserves are crucial for business, it is hard for the sovereign to demand large levies on the grounds that it will have little impact on business. The idea is related to Calomiris and Kahn (1991) who take the asset structure of the bank as given and argue the liability structure emerged to control agency problems. Here, I argue that the liability structure arose from the payment function and the assets were chosen to minimize problems of agency.
Richards (1929, pp. 35–36) suggests that this incident, and the general insecurity to property accompanying the Civil War, was the impetus for a system of private banking. Merchants sought places where their excess specie could be safely deposited without attracting the unwanted attention of the unscrupulous sovereign (or parliament). As a result, a number of goldsmiths developed into deposit bankers, and started making loans. Richards does not explain why the goldsmiths offered better protection than the Mint. The above argument that a hoard dispersed among many goldsmiths, and in large part further dispersed to borrowers, is both harder to seize and elicits more opposition when seized, provides the missing link in the explanation.

I do not mean to suggest that there were no abuses possible in overdrafts. All I am arguing is that given the abysmal level of accounting, monitoring, control, and enforcement of property rights at that time, banks could not be trusted to sit on too much reserves. Thus overdrafts not only amortized fixed costs of liquidity provision but also were a second-best arrangement for ensuring that the depositors’ money was safe.

**Toward an Incomplete Contract View of Banks**

Clearly, banks emerged, in part, to resolve contracting problems—as suggested above, asymmetric information about the value of coins. Similar contracting problems have been studied before in the literature, and the bank usually emerges as the optimal contract, or “nexus of contracts,” that resolves these problems. However, the historical discussion suggests that contractual incompleteness and the limited security of property rights were critical to determining the institutional structure of the bank. This is where the incomplete contract approach to banking is valuable, and deserving of further study. Let me elaborate with some examples. My intent, however, is not to explore all the possible applications of incomplete contracts to banking, nor to present a comprehensive theory.

First, it is useful to distinguish different forms of incompleteness of contracts, something the literature rarely does. Sometimes contracts are not enforceable because the authorities do not respect property rights or because it is too costly to use the legal system to verify outcomes. This is external incompleteness where the characteristics of the external legal and property rights environment prevent detailed contracting. For example, in advanced legal systems where judges are relatively incorruptible, detailed explicit priority rules for repaying different classes of debt can be enforced through a relatively transparent bankruptcy proceeding. In more primitive system, perhaps only the crude right to liquidate assets is enforced with any degree of transparency for anyone other than the most powerful creditors. A variety of institutions will then spring up to substitute for external contract incompleteness.

Consider the rise of the bank. Given the limited enforcement of property rights, the historical evidence suggests the institutional form of the bank arose to give appropriate amounts of power to the various stakeholders. For instance, medieval laws allowed creditors to seize assets of delinquent borrowers. However, given the limited

extent of markets in those times the assets were probably illiquid and worthless in the hands of an ordinary lender. Because the medieval banker aggregated the information, financial, trading, and political power of his depositors, he had a greater ability to recover loans from borrowers (Kindleberger 1993, p. 47). But if the banker was so powerful, what made him repay his depositors? Myers and Rajan (1998) argue that because the banker had to maintain extremely liquid reserves to continue his business, depositors could be confident that they could recover their money from him. The intrinsic liquidity of the coins the banker held as reserve could enable depositors to confidently threaten liquidation without having the powers or the connections the banker had. Thus the institutional form of the bank allowed a distribution of power, even absent sophisticated legal enforcement, that permitted the flow of credit.

There are other forms of incompleteness. It may simply be very hard for a lender and borrower to contract on eventualities because they are too hard to describe or visualize in requisite detail. This is *intrinsic* incompleteness. For example, it is very hard to contract on a borrower's creditworthiness, since creditworthiness involves many judgmental factors and character issues that a loan officer cannot communicate in verifiable detail to a court. It is also very hard to anticipate all the factors that will affect a borrower's creditworthiness. The last form of incompleteness I would like to focus on is what I call *deliberate* incompleteness. Here contracts are not written in all the detail possible simply because doing so would lead to worse outcomes than leaving contracts incomplete.  

When contracts are incomplete, there is bound to be bargaining after outcomes are realized. The point of institutional design is to structure institutions such that the bargaining power goes to the right party in the right circumstance. Unlike the mechanism design approach that has dominated the literature so far, the incomplete contract approach focusses attention on bargaining and the institutional structure that shapes the rights and powers each party has in the bargaining.

To see the value this approach has, consider how it explains some institutional details about the bank. I have argued that the bank offered liquidity to both depositors and borrowers. Why did it make only the former's claim explicit while leaving the latter's claim implicit—the bank could refuse a demand for an overdraft but not a demand for repayment of a deposit. The answer must be that given the difficulties of enforcement (external incompleteness) and given their individual weakness, depositors needed a cast-iron commitment of repayment from the bank else they would not deposit. By contrast, it would be very hard (even today) to specify who would be deserving of funding (intrinsic incompleteness). Since it would be socially inefficient to fund under a variety of circumstances, it probably made more sense to leave the right to be funded implicit and to be negotiated. Using the jargon, liquidity demand for credit was observable but not verifiable, so it could not be contracted upon (or only contracted upon in an overly crude way with attendant costs). The bank, however, had a reputation to maintain for being accommodating. Therefore, the bank could offer

8. Of course, in a world where all state-contingent outcomes could be contracted on, this would not make sense. But if some outcomes cannot be contracted on, it may make sense not to contract on others also (see, for example, Boot, Greenbaum, and Thakor 1993).
implicit promises of credit to potential borrowers, and the latter could be confidant that the promises would largely be honored. At the same time, the bank could renege on the promise if its circumstances were too onerous. Thus the bank could optimally substitute between reputational and monetary capital in ways the market could not [see Boot, Greenbaum, and Thakor (1993) for a theoretical exposition of this point]. This could explain why the potential borrower’s claim was deliberately left incomplete.

What distinguishes institutions from arm’s length contracts in the incomplete contract approach is then that institutions overlay the contracts the sign with noncontractual attributes, many of them related to specific investment that its clients, its employees, or the institution as a whole have made. For instance, in the above discussion, the bank enjoys a valuable reputation that stems from past specific investment in reputation building that would not be available to an anonymous participant in the market.

Before concluding, a caveat is in order. The danger in postulating a particular kind of incompleteness to explain a particular institutional feature is that one can obtain the same vacuous theory that the asymmetric information literature sometimes throws up—the folk theorem that for every phenomenon, there exists an asymmetry of information that explains it. As always, the best way to counter this criticism is to make multiple predictions on the basis of sparse assumptions so that the theory explains more than the phenomenon it was built to explain.

Summary

The purpose of the historical discussion above is not only to highlight the functions that banks provided, but also to suggest that the importance of banks depended to a large extent on the inadequacy of the legal and informational environment. Banks were second-best institutional solutions to the problem of payment, savings, and credit in an environment of incomplete contracts.

2. REGULATIONS

I have argued that money changers combined deposit taking with overdrafts partly for standard economic reasons—to spread the fixed costs of maintaining liquidity and to utilize information gained about depositors. The products were also jointly offered because the combination offered synergies when property rights were inadequately enforced—the money changer aggregated the power of his depositors to extract repayment from recalcitrant borrowers, he dispersed liquid funds among borrowers essentially reducing his temptation to commit fraud on the depositors, as well as the government’s temptation to “borrow” the funds, and he set up the deposit contract so that depositors could recover their money from him.

In addition to the private response by the money changers themselves to the inadequacy of contracts, there was a public response by medieval authorities. It is hard to tell whether the regulations they implemented were motivated by the public interest or self-interest. Nevertheless, the pattern set by early regulations has been followed since, hence they are worthy of examination. The following is not, however, intended
to be comprehensive [see Kane (1997) and Benston and Kaufman (1994, 1996) for excellent surveys on the issues].

As evidenced by the frequent attempts to set up municipal or public banks, an important reason for regulating banks was to limit the risk to the payment system. One can surmise that disruptions in credit were thought of as less important because the municipal banks were enjoined from offering credit. Another reason to regulate banks was, of course, to tax them. Let us now consider the kinds of regulations imposed in more detail and the way they affected the functioning of the banks.

Regulations to Penalize Fraud

As suggested above, and as borne out by a number of studies of banks even in the modern era (Calomiris and Kahn 1991), the major cause of bank failures prior to the introduction of explicit government deposit insurance was dishonest management. Early regulators tried to protect against banker fraud by forcing them to put up monetary capital as collateral. In the typical laws regulating banking that were passed in the late thirteenth century, money changers were required to “capitalize” their banks by depositing a sum of money with the city government—this amounted to a thousand marks of silver in Barcelona in 1301. Only then were they allowed to spread a tapestry bearing the shield of the city on their table indicating that their office was guaranteed. Those who did not pay the fee had to leave the top of their tables bare as a warning to prospective depositors. Furthermore, if the money changer failed, his property was held liable to the claims of his creditors.

Given the temptations the money changer faced, the threat of seizure of monetary capital was not always adequate. So regulations, typically passed in times of distress, attempted to secure his human capital as collateral. Penalties extended to the money changer’s future ability to earn, his social relationships, and his person. An ordinance in Barcelona passed in 1300 stated (Usher 1943, p. 239)

No money changer who may hereafter fail, and none who has recently or in times past failed, shall again keep a bank or hold any office under the crown. Any such shall be proclaimed bankrupt and disgraced by public crier throughout the city and also in the place he had his office. Until he shall have satisfied all demands, he shall be detained on a diet of bread and water.

If no settlement was made after the money changer failed, the money changer was beheaded and his property sold to repay creditors. Furthermore, since money changers whose affairs became seriously compromised were liable to flee before the facts became known, “prelates, barons, or knights” in whose territory they sought refuge were required on request of the crown to “capture, hang, and behead” the bankrupt money changer. These were severe penalties, but the city governments thought them warranted given the importance of the function and the temptation of fraud.

Regulation to Provide Positive Incentives to Not Commit Fraud

Of course, many of these penalties could not be levied if the banker fled to enemy territory, beyond the reach of the authorities. A more humane way to give bankers the right incentives is to give them economic rents if they conduct their business appro-
propriately. Thus regulators often restricted competition among bankers, thereby creating franchise value in some part of the banking process. For instance, regulators allowed the privilege of offering demand deposits and payment services to only a few banks, thus creating supernormal profits in deposit taking, and, in effect, capitalizing the few. Since the banker had to stay in business in order to collect the rents, depositors would have been more likely to trust him with liquid assets than they would if the industry were competitive.

Also, by conferring regulatory privilege on a few, regulators limited the number of entities that had to be supervised. To the extent that even regulators have a limited capacity, this may have been seen as a benefit to restricting competition.

**Regulation as Taxation**

A more cynical, and often more accurate, view of regulation is that it is motivated by the authorities’ need for finance. By offering banks a monopoly (of note issue or deposit taking), and then skimming some of the profits by demanding (or directing) low-interest loans from the banks, the authorities could pretend to ensure the safety of the public’s investment while effectively taxing them (for example, see Benston and Kaufman 1996 and Kroszner 1997). It is not particularly surprising that many central banks have come into existence at a time when governments have been in dire need for finance (Goodhart 1985). Hammond (1957, pp. 152–59) has a particularly entertaining description of how rival political parties in New York pushed through charters for their “own” bank when they were in power.

Nevertheless, even if the government creation of banking monopolies was motivated by self-interest rather than the public interest, it served to reduce the public’s uncertainty. The government could now institutionalize its tax on the system. So long as the government did not become too greedy, it is even possible that the effectiveness of the banking system could have been enhanced. To understand why, we have to go back to our idea that banks play a role when contracts are incomplete.

**Regulations as Strengthening Implicit Contracts**

I have argued that banks are in the best position to offer credit at short notice to clients. The problem often is that needy bank clients will generate the rents to repay the credit in the long run but cannot commit to doing so because of the incompleteness of contracts (see Mayer 1988). For instance, start-up firms need substantial amounts of credit when they are young. From the bank’s perspective, the interest rate that the bank would have to charge on the initial loan in order to offset the short-run costs of monitoring the firm would create severe agency distortions. It is better that the bank recoup its up-front costs by spreading them over the future business it does with the firm. Unfortunately, in competitive banking markets, the firm may find other banks more attractive if it survives beyond the initial years. Unless the initial lender has a commitment from the borrower to use it for all future borrowing—and these contracts are hard to write and even harder to enforce—the initial lender is left with a loss. Thus the start-up does not get financed. A similar problem exists for distressed firms.
The virtue of a more concentrated banking system is that a bank is more likely to finance young and distressed firms because it knows that it will be the only source of funds the firm will have access to for a long time. So the bank can internalize the long-run benefits of assisting the firm. Petersen and Rajan (1995) find evidence in the United States that young firms get more financing in areas where banking is more concentrated, and pay a lower rate than young firms in more competitive banking markets. Thus regulations limiting competition not only make the banking system more stable (again provided the government does not make undue demands) but also make more effective the implicit contracts in which banks specialize.

Of course, a bank recoups the cost of the initial subsidy it offers by charging higher-than-competitive rates when firms are older. This suggests that there are also standard deadweight costs of offering banks monopolies which must be offset against the enhanced noncontractible possibilities this leads to. For a country at an early stage of development—when credit availability is important—it may be beneficial for banks to have some market power. But at a later stage of development, the price of credit is more important, and bank market power conferred by regulation can retard growth.

This suggests that there is an optimal time for a social planner to introduce competition in a banking system. Of course, before planners in nascent transition economies decide to confer privileges on banks, they should remember that banks have a very strong incentive to oppose curtailment of these privileges, and optimal deregulation will, typically, not take place. Furthermore, it may well be that the incompleteness that prevents long-term contracting between the bank and the borrower could be addressed by the government in a less interventionist way than giving bank monopolies. For instance, it may be sufficient to allow banks to hold equity stakes in firms.

**Summary**

Banks arose to provide liquidity and safety to customers at a time when information and financial knowledge was scanty, property rights poorly enforced, and might was often right. The combination of lending and deposit taking enhanced the kinds of transactions that could take place in this environment. Regulations typically increased the liability of bankers above those of other merchants. This increased the credibility of bankers because they could offer their wealth and human capital as collateral (Winton 1993). Regulators also restricted competition. While not necessarily motivated by the public good, such restrictions probably made banks safer by giving them franchise value (Keeley 1990), and enhanced transaction possibilities between banks and their clients.

3. **WHAT CHANGED?**

Obviously, much has changed over the last six hundred years. Most of the change has really been concentrated in the last century. Even in this narrower period, I would
like to emphasize only three important areas of change: technology, information availability, and the property rights environment.

**Technology**

The advances in computing and communications technology have been tremendous. Fundamentally, the cost of tallying transactions and keeping track of positions has been reduced to next to nothing. While historically, months and even years could pass before a money changer’s accounts were fully tallied, today, mutual funds with enormously complicated positions post their net asset value every day. Similarly, as long as there were substantial costs of transacting at a distance, banks enjoyed cozy local monopolies. This is no longer the case. Retail customers can undertake banking transactions on phone lines or ATMs from distant banks without ever setting foot in them.

**Information Availability**

The ability to communicate information to a wide audience which has the capacity to process and digest it is a necessary condition for democratizing the availability of information. It is also necessary that agents be willing, and able, to produce and sell the information over the communication lines. A variety of information producing agencies such as rating agencies (Moodys, S&P, etc.), data gatherers (Lexis/Nexis, Datastream, etc.), and disseminators (Reuters, Bloomberg, etc.) have emerged to do this.

There has been substantial variability on this dimension across countries. The paucity of empirical corporate finance studies in continental Europe to date is not just because of the strong rationalist tradition in those countries but also because of the reluctance of companies to disclose data. Corporate accounting standards in countries with an Anglo-American legal tradition have been much higher than in countries with a French or Germanic tradition (Laporta et al. 1996) suggesting that corporations have not been forced by either the public or the authorities to share information.

**The Property Rights Environment**

Institutional checks and balances on governments have reduced the extent to which blatant government extortion has to be feared. Property rights have become much better defined and contracts are more easily enforced. Judicial systems have become less corrupt as they have come under the glare of public opinion. Especially important, the democratization of information coupled with the democratization of political power has given the public investor strong weapons against corrupt management.  

---

9. This was brought home to me in a conversation with a prominent professional shareholder rights activist who suggested that she managed to change the composition of boards of companies in the Fortune 500 list simply by threatening to take out a full-page advertisement in the Wall Street Journal with their names prominently featured. Of course, such activists have gained power only recently, and there are still many economies where such activism would be ignored or, worse, suppressed legally or illegally.
Summary

The changes I have identified have both increased the sense of security in property rights and permitted third-party monitoring, thus reducing external incompleteness. The ability to write contingent contracts has also increased because of the increasing number of hard-to-influence market-wide indices that are now computable. For example, it is now possible to tie contracts to market-wide interest rates, credit spreads, oil prices, etc. Thus intrinsic incompleteness has also decreased. Since I have argued that banks essentially were institutional solutions to situations of incomplete contracts and inadequate property rights, it is reasonable to ask if the decline in banking documented earlier is a result of these changes. This is what I explore next.

4. HOW HAS THE BUSINESS OF BANKING CHANGED?

I start by asking if the functions that banks historically performed—of liquidity provision and safety—have become less important.

Liquidity Provision Revisited

The wide availability of information and improvements in processing technology have improved financial markets to the point that large firms can raise money at very short notice from the public commercial paper markets. As a result, the ratio of non-bank commercial paper issuances to Commercial and Industrial (C&I) loans made by banks rose from 10 percent in 1958 to 75 percent in 1990 (Gorton and Pennachi 1995). Despite these numbers, it appears that commercial banks still play a role in guaranteeing liquidity.

Benveniste, Singh, and Wilhelm (1993) find that when Drexel Burnham Lambert filed for bankruptcy in 1990, the share prices of major money center banks went up by an astonishing 7 percent. By contrast, the share prices of competing investment banks went up only by 1.6 percent. Clearly, Drexel was more of a competitor to commercial banks than to investment banks. But what exactly did Drexel do? Essentially, Drexel’s Michael Milken set up a network of junk bond buyers who accepted pretty much any deal he sponsored. In return, he exercised control over the borrower to make sure they did not lose out. Furthermore, if a particular issue underperformed, he made it up to the buyer in future transactions. With relationships in place with buyers and thus placement power assured, he could guarantee borrowers short-notice finance—exactly what I argue a bank does. In fact, Drexel’s letter assuring a company board under hostile attack that Drexel was “highly confident” of raising finance for the takeover became virtually a loan commitment, albeit for a very large sum of money. No wonder bank stockholders rejoiced at Drexel’s demise!

Since there are fixed costs in arranging for liquidity even today—the fixed costs of maintaining reserves, maintaining borrowing or syndication links with other institutions, and enduring the costs of being supervised in order to get access to the discount window—which can be amortized over multiple and diverse demands for liquidity,
we would expect the bank to have a comparative advantage in issuing guarantees that require coming up with funds at short notice. Its advantage will be especially strong when a firm’s financial situation is novel, complicated, or distressed. In these situations, the bank’s institutional ability to enter into longer-term relationships (and thus its ability to mitigate incompleteness) will give it a comparative advantage over more market-based solutions.

But if banks still have a role to play in liquidity provision, what explains the rise of the commercial paper market?

**Bank Opacity and Securitization**

Banks have not withdrawn from the commercial paper market. Instead, banks have unbundled what historically went together—the initial loan and the commitment to provide support in times of distress. Now they allow investors to lend to the firm directly through commercial paper while providing back up lines of credit or letters of credit to assure the commercial paper investors that they will get their money back.

But why have banks moved away from funding loans directly to simply guaranteeing liquidity? The answer, I think partly lies in the dramatic change in what used to be the bank’s primary assets: loans. Bank loans have, historically, been illiquid. The originating bank that has the relationship knows more than the potential buyer about the borrower’s true credit risk. So a potential buyer of bank loans faces the risk of getting only “lemons,” which may be one reason the market has been illiquid. Another reason has been that bank loans typically involve more complicated covenants than public debt, and are usually smaller in size. The computational costs of keeping track of title, payments, and covenant violations may not have justified the sale.

All this has changed. Banks have started selling their loans. The loan sales market has grown tremendously; from approximately $26.7 billion in the second quarter of 1983 to $290.9 billion in the third quarter of 1989 (loan sales declined somewhat in the early 1990s as loans financing mergers and acquisitions declined). What is also notable is that the majority of loans sold changed over this period from investment grade to being noninvestment grade (Gorton and Pennachi 1995). A number of factors are responsible for the greater fungibility of bank loans. Greater information availability about firms has reduced the asymmetric information about borrowers, greater computational power has reduced the costs of slicing and dicing up loans into uniform, saleable pieces, and the loan sales are structured so that the originator retains a stake. Moreover, increased volume can improve liquidity. Buyers do not inspect what they buy very closely. Instead, they trust sellers. The reason they can do so is that the greater integration of markets has increased the frequency of transactions any single player undertakes. Reputation not only becomes easier to build, but also more important to maintain as banks fund loans through their placing power rather than their balance sheets.

What is interesting is that the ability to sell loans, paradoxically, makes it very hard for banks to fund high-quality loans on their balance sheet. Since the bank has the option to sell its loan, and bank balance sheets are typically opaque to the investors,
long-term investors in the bank who cannot frequently reprice their claims charge a hefty premium for funding the bank. Furthermore, the rate of interest demanded by these investors does not respond to the marginal loan made by the bank because investors do not see the loan, and have no idea how long it will be on the balance sheet.

To make this point clear, consider a bank that is A rated and borrows at 12 percent. In a perfect Modigliani-Miller world, it can make a new loan to a AAA customer at 10 percent and still make a profit. The rationale is that the bank’s lenders, seeing the new safe loan, will adjust downward their estimate of the bank’s credit risk, and demand a lower rate. In a perfect world, the downward adjustment on the bank’s average cost of funds will be just enough to make the loan worthwhile. But if bank balance sheets are opaque, their cost of capital does not adjust downward to the marginal high-quality loan because investors do not know whether, and for how long, it is there. A bank’s cost of capital for a new loan is thus not independent of its existing credit rating, unlike a Modigliani-Miller world. A bank that wants to profitably lend to high-quality credits has to either bolster its capital so that its own credit quality improves, or find a convincing way to commit to the market that it will keep only high-quality loans on its balance sheet—for instance, by changing its clientele over time. The transition may be long and very costly.

This then explains the puzzle of why banks guarantee the commercial paper rather than making the loan directly. As argued earlier, banks have a special advantage in providing liquidity when markets sour on a firm, and the firm finds it hard to pledge enough of its future rents to get credit. Rather than lending to a firm and keeping the loan at high cost on its balance sheet, it makes sense for the bank to lend only on a contingent basis—when all other sources of funds dry up and the firm is a high risk. In more normal times, nonbank investors fund the commercial paper directly. The virtue of this is that they know what they are investing in and charge a rate appropriate to the risk of the firm rather than a rate commensurate with the blind-pool average risk of the lower-credit-quality bank. This explanation could also account for the vast amount of securitization of ordinary loans that is taking place (see Myers and Rajan 1998).

One could ask why an institution is still needed to provide liquidity to the commercial paper. Could assurances of liquidity not be obtained from the market? The earlier analysis suggests three answers. First, given the incompleteness of contracts even today, a blanket assurance of liquidity is probably all that the market can offer. By contrast, an institution can offer a contingent assurance of liquidity where the institution commits with its reputation not to renege unless the contingencies are really adverse—in other words, it leaves the contract deliberately incomplete and effectively completes it through its own reputation (see Boot, Greenbaum, and Thakor 1993). Second, a potential borrower may not be able to assure a market investor of enough compensation for the liquidity if they were to write a single contract, while the borrower may be able to compensate an institution because they are tied together over time and over different products. Finally, institutions may provide liquidity in more innovative ways.

10. Of course, most bankers go too far in the other direction and refuse to recognize that their cost of capital will adjust downward if they keep making loans to firms that are better rated than they are.
This last point is related to Merton (1995) who argues that financial innovations (such as commodity swaps or credit derivatives) are first developed by institutions and then, when the financial contracts are well understood, are taken up by the market. Financial institutions then move on to new products. Thus there is a continuous spiral of innovation. He does not explain why financial innovations are better produced by financial institutions rather than directly by exchanges. The answer, I think, again lies in incompleteness. Innovative financial contracts are typically intrinsically incomplete in many ways when they are first introduced. Payments or responsibilities have not been fully spelled out for many possible situations, partly because the situations themselves have not been anticipated. A trial period is necessary where the contract has to be tried out in real world situations, and the appropriate contractual features to deal with initially unforeseen contingencies have to be developed. The firms with which an institution has relationships form the ideal testing ground. As above, institutions manage incompleteness better by allowing reputational capital to fill in the holes that explicit contracts cannot fill.

To summarize, it appears that the provision of liquidity is still an important function of financial institutions. One could even argue that with the greater informational integration of financial markets, it has become even more important. An incorrect rumor in one financial market can quickly spread to another, through contracts contingent on public information [see Diamond (1991) for the theory]. This may cause a run on a firm’s paper unless it has deliberately incomplete institutional relationships that help staunch the run. Finally, even though there may be a secular trend in contracts becoming more complete, regions of incompleteness are large enough that institutions can simply migrate to new areas where they have a comparative advantage in liquidity provision. Let us now ask about the function of providing safety for invested money.

The Provision of Safety

Investors now have access to many other savings vehicles than demand deposits. It is now possible for them to invest in secure money market mutual funds that are constrained through regulatory oversight to invest in a very narrow set of securities. The liquid assets that are held in the fund are monitored on a daily basis, and are protected by a variety of institutional arrangements such as third-party custodians from fraud. Furthermore, these institutions have partial access to the payment system, and only the political power of banks prevents them from having full access.

At the same time as other safe investment vehicles have appeared, the widespread availability of real-time information about individuals and firms has reduced the informational link between deposits and overdrafts. Individuals have widely available credit rating histories that are far more informative than the time series of their deposit account balance. They are no longer compelled to borrow from the bank they deposit

11. Adequate custodial arrangements are not a trivial institutional development. As recently as 1983, Chase Manhattan suffered a loss of hundreds of millions of dollars when First Drysdale Securities, a swaps dealer, defaulted on securities it had promised but did not have.
with. Perhaps a small-town bank knows more from monitoring the checks its small customer writes than do outsiders, but a number of agencies like Dun and Bradstreet compile data from a variety of sources on a firm’s payment history that far exceed what a large bank (which may be one of twenty lenders) has private access to.

Since the informational link between deposits and loans is no longer strong both for individuals and for firms, 100 percent reserve banking is now a possibility. So if deposit insurance and privileged access to the payment system is done away with (see below), would banks continue offering deposits?

I think the answer has to be an empirical one since the theory makes ambiguous predictions. Certainly, depositors would migrate to money market funds if they were interested only in safety. But money market funds also sacrifice substantial returns by staying fully invested in liquid assets and not “using” the liquidity of those assets to provide liquidity to others [see Kashyap, Rajan, and Stein (1998)]. If banks were to compete directly against money market funds, they could earn enough from working their liquid assets harder to pay higher rates to compensate for their higher riskiness. There is a risk-return trade-off between institutional forms, and it is ambiguous which way it goes.

5. THE ROLE FOR REGULATION

As I have argued, early regulation was intended by the government to protect the critically important payment system and to secure credit for the government. Do these rationales exist any more?

Why Regulate?

Some argue that the payment system is still critically important and participation in it should be the criterion for regulation. While the importance of the payment system cannot be questioned, it is possible to protect it from the risk of failure of even a large bank by using a fully collateralized real-time settlement system. It is hard to believe that the costs of such a system (for instance, in terms of reduced liquidity or the costs of posting collateral) outweigh the benefits of doing away with supervision. Furthermore, even if intra-day credit is deemed important, there is no reason why private arrangements to offer it, or to share and limit the risk, cannot be devised [see Hoenig (1996) for an excellent discussion].

The effect of insulating the payment system from the credit risk of specific institutions reduces the need to ensure the safety of institutions that have access to the payment system. This then eliminates the need to restrict access to a privileged, supervised few. Instead, any institutions that can post high-quality collateral or obtain private guarantees should have access. Certainly, there is no reason to bar money market funds from having full access to the payment system.

If safe institutions such as money market funds have full access to the payment system, any economic rationale for insuring deposits at banks evaporates. Some (see
Kaufman and Benston 1996, for example) conclude that deposit insurance cannot be eliminated for political reasons because of the opposition of small banks. This may well be true. Also, even if a bank does not have explicit insurance, it is likely that a large bank dealing with many retail depositors would have implicit insurance whatever the ex ante declarations by the government. The political pressure on democratic governments to bail out the foolishly greedy, so long as they are numerous, has been demonstrated time and again, most recently with Ponzi schemes in Eastern Europe.

Once deposit insurance exists, whatever the reason, Benston and Kaufman (1996) argue the role for regulation is to make sure that the distortionary incentives offered by deposit insurance are counterbalanced by supervision and prompt corrective action (which consists of imposing increasing strictures on a bank’s ability to operate as its capital level falls). Thus one, largely anachronistic, regulatory privilege breeds other, more intrusive regulations intended to minimize the use of the privilege. But is there any good reason left to regulate?

**Liquidity Crisis**

Thus far, we have looked at interventions that were intended to ensure the safety of the payment system. Another rationale for intervention is to ensure the flow of credit. The incomplete contract approach suggests that the costs of financial distress are often large for a bank, and also for small and growing corporations who are its clients. A bank may have a special ability to offer credit to customers because it can get better repaid. There are two components to this ability—reputation and trust that enables non-contractible transactions to go through, and a long-term relationship in which both parties have made specific investments. If a bank is hit by a severe shock, it faces a potential end-game situation which reduces its reputational commitment as well as the expected life of its relationships. Thus the bank’s ability to offer commitments and recover loans falls off sharply as it approaches distress, a cost over and above any fragility on the liability side. Bank failure may have a multiplier effect in the local economy, because other banks may not have the relationships or specific knowledge to step up the flow of credit to firms. The effect of a bank failure will be particularly serious for small and medium-sized firms who have fewer relationships (see Petersen and Rajan 1994). A crude measure of the size of these costs emerges from a study by Slovin, Sushka, and Polonchek (1993). They find that in 1984, client firms of Continental Illinois Bank incurred average abnormal stock returns of $-4.2$ percent during the bank’s impending insolvency and $+2$ percent in response to announcements of the government rescue.

12. Kaufman (1994) presents evidence suggesting that the economic rationale for deposit insurance—that it prevents contagious bank runs—is not borne out by the data. Bank runs do not seem to be based on random events, but rather on specific information. However, as Diamond (1991) shows, even information-based runs can destroy social value. Whether disciplinary effects of information-based runs outweigh the ex post social costs is an empirical question.

Of course, the existence of these costs does not imply regulatory intervention is warranted. The rationale for regulatory intervention is that by standing by the bank the government enables it to continue in business. Private parties cannot offer such guarantees because they do not internalize the spill-overs and, more important, they do not have the powers of the long-term relationship the government has to recover the costs of the guarantee from the bank. Of course, even if the government steps in to rescue at low cost, it reduces the ex ante incentives that banks have to stay out of trouble. Unless government intervention is accompanied by stringent penalties to bank management (it may not be possible for them to be severe enough because of the possibility of destroying the relationships that the government is trying to rescue) the ex ante distortions to incentives may outweigh the ex post benefits.

This is not a blanket endorsement of ex post intervention. Certainly intervention when there is a systemic loss of confidence is more warranted than when a single bank is distressed. Intervention when a bank’s clients have no recourse makes more sense than when they have alternatives. In a general panic, this suggests the government should step up immediately to enable banks to continue transacting, but then quickly conduct a serious audit of those that need to be rescued and those that do not. Also, rescue may well take the form of encouraging mergers (provided there are enough healthy banks) since a healthy bank can probably extract more from the failed bank’s relationships than can the government.

Who Should Be Supervised?

In summary, once the payment system is immunized from any individual institution’s risk of failure, it is less clear that depositary institutions are special and should be supervised. As one extreme, it would appear that every institution that has implicit or explicit claims on the public purse should be supervised, ranging from finance companies that deal with the public to investment banks that have relationships with small firms. This is obviously too much. Supervision may breed more distortions. For instance, in the case of retail financial institutions, the retail customer’s implicit claims for a bailout will be strengthened if the institution he deals with is supervised for safety and soundness. A better route is to cut any retail financial institution’s anticipated links to the public purse whenever feasible.

For wholesale financial institutions who are in the business of providing relationship credit, some light supervision may be warranted to ensure that they do not collectively load up on the same risk, be it interest rate exposure, foreign exchange exposure, or credit exposure to a particular sector or country. Even here, full use should be made of market-based mechanisms to alert supervisory authorities of problems [see Wall (1989) for an ingenious suggestion, and Flannery (1997) for an excellent survey of the efficacy of market-based methods]. At the same time, there seems

14. Why not rescue industrial firms also? I would argue that with industrial firms, much of the incompleteness in contracts is internalized through vertical integration. With financial intermediaries, much of the incompleteness continues to reside in the transactions it has with customers, hence the importance of reputation and long-term relationships.
little reason to restrict regulatory privileges such as access to the discount window only to banks who, by accident, hold a charter. The discount window should be opened only in times of systematic failure, but be freely available to all institutions that are creditworthy (where creditworthiness is measured in the long run, and from the government’s perspective) and who provide substantial amounts of liquidity insurance (since a crisis at an institution will spread to clients only if they need substantial amounts of refinancing in the short run). Of course, there is potential misuse or favoritism in the use of the discount window, but it can be targeted much more effectively than open market operations [see Kaufman (1991) for a different view].

In summary, once the payment system is immunized from the idiosyncratic risk of participants, the only economic reason for regulatory intervention is if there is a systemic crisis in institutions specializing in relationship credit. These institutions may, or may not, be what have traditionally been called banks.

*The Changing Institutional Form*

If regulation becomes more even-handed and deposit-taking institutions no longer have special regulatory privileges, will liquidity providers continue to be depository institutions? Since large institutions now offer liquidity through such a large number of off-balance-sheet products to such a large variety of customers, offering deposits may have no additional diversification benefits. If deposits bring no regulatory privilege and little diversification benefit, it is reasonable to expect that some liquidity providers will not offer deposits. Already, some of the largest banks deemphasize transactions deposits (see Kashyap and Stein 1997).

Additionally, since the product mix will be ever changing to meet the needs of customers, the positions of such an institution will be hard to track, even if the institution were regularly audited (by regulators or private third parties). Thus unlike the money changers of old who could rely their own product mix and on regulators to assure customers that they would be around to provide liquidity, such mechanisms will not be available to the modern liquidity provider.

These institutions, however, have to provide assurances that they are rock-solid since their ability to guarantee liquidity does not stem from regulatory privilege. Who would buy a guarantee from an institution that has a high chance of going bust? Capital can go only so far in producing assurances—recall that Nick Leeson managed to wipe out Barings’ capital in a few unsupervised days, long before any regulator got wind of it. Critical, then, for the modern liquidity provider are good risk control systems. By this, I do not just refer to the models the institution uses to assess risk and allocate capital, but also the system of employee compensation and discipline, the internal organizational structure (see Kahn 1992; Kroszner and Rajan 1997), the internal audit system, the governance exerted by the board, and even the corporate culture. If there is an institutional core to the modern financial firm providing liquidity service, it is these specific investments in internal organizational control and coordination, rather than any specific products it offers such as deposits.
A Functional Approach?

One cannot conclude a discussion of regulation without mentioning the “functional approach.” Robert Merton, in a series of articles (for example, Merton 1995), has proposed a functional approach toward understanding and regulating financial institutions. Essentially, he argues that the financial institution should be viewed in terms of the functions it provides rather than its institutional appellation. There is much to commend in this and the approach has revolutionized thinking among many regulators (and academics). The one major problem, however, is that it uses too much of a contractual approach to analyzing institutions. If the institutional core is noncontractual, then viewing the institution merely in terms of the separable functions it provides (or contracts it signs) may miss what is essential to the institution. In fact, the incomplete contract approach argues that it is precisely the interaction between functions and the noncontractual core that makes the institution better than arm’s length contractual solutions. A functional approach to regulating institutions may assume a degree of separability that simply does not exist at present in this world of incomplete contracts.

6. ARE BANKS DYING?

Let us go back to the question that started the paper: Are banks dying? If the question refers to those institutions with a bank charter, then the answer must be no, especially when we take account of nontraditional activities. When noninterest bank income is capitalized and added to bank assets (and foreign bank loans are accounted for), Boyd and Gertler (1994) argue that average bank share of intermediated assets is stable at about 40 percent between 1955 and 1993.

If the question refers to the institutional form, we have to be more careful. There are certainly an enormous number of small community banks that still offer very traditional products. But the bulk of the assets in the banking sector are in large banks, where the traditional products form a decreasing share of operations. Some of these banks like Bankers Trust and J. P. Morgan are wholesale liquidity providers who are more similar to investment banks like Goldman Sachs than to their community bank cousins. Others like Citibank have an increasing retail focus and offer investment and transactions services to the retail customer, much like the nonbank discount brokerage, Charles Schwab. So the traditional bank has given way to institutions that focus on different segments of the market—for instance, wholesale customers or retail customers.

If the question is posed in other countries than the United States, such as in Continental Europe where regulatory privilege and suppression of public information about firms still give banks some rents, the answer is “Not yet.” But with deregulation and the integration of European financial markets, there is no reason why the developments in the United States will not visit banks in those countries, too.

Finally, if the question is meant to ask if in an ideal, apolitical world, some bank
regulators should worry about job security, the answer is yes. But, of course, that is a Chicago School utopia that we are unlikely to see.

LITERATURE CITED


Lummer, Scott, and John McConnell. “Further Evidence on the Bank Lending Process and the


