JOURNAL OF FINANCIAL INTERMEDIATION 4, 158–187 (1995)

Delegated Monitoring and Bank Structure in a Finite Economy*

ANDREW WINTON

Finance Department, Kellogg Graduate School of Management, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208

Received January 13, 1994

When banks act as delegated monitors of borrowing firms in a finite economy, two factors help banks dominate direct lending: portfolio diversification, which increases with bank size, and bank capitalization, which diminishes with size. With free entry into banking, intermediated equilibria are possible even when direct lending cannot overcome autarky. There are usually multiple intermediated equilibria; these may not be Pareto-ranked by bank size, since smaller banks are better capitalized and may Pareto-dominate larger banks. Even when one large bank would be most efficient, assigning a monopoly bank charter to coordinate beliefs on the single bank equilibrium may be unattractive: in some cases, a monopoly bank cannot overcome autarky even though free-entry banking can, and in other cases, the monopoly bank reduces production from the direct lending level. *Journal of Economic Literature* Classification Numbers: G21, L13, O16.

1. INTRODUCTION

As banking systems around the world undergo stress and change, it would be helpful to have theories that examine bank behavior in a context that captures why banks are useful in the first place. Although there has been much research on how banks can help overcome capital market imperfections, much of this research focuses on a single bank in isolation.

* This paper is derived from Chapter 3 of my Ph.D. dissertation, which was supported by a Unisys Fellowship. Earlier versions were entitled "Banks as Delegated Monitors: Industrial Structure in a Finite Economy." I am grateful to Franklin Allen, Arnoud Boot, Francisco Delgado, Peter DeMarzo, Mike Fishman, Gary Gorton, Stuart Greenbaum, Joe Haubrich, Rich Kihlstrom, Debbie Lucas, Tony Santomero, and two anonymous referees for their help and advice. All mistakes remain my own responsibility.

1042-9573/95 $6.00

Similarly, although there has been much research on the impact of regulation on equilibrium bank behavior, much of this research does not model how banks add value.[1] My paper combines these two issues, examining general equilibrium in an economy where banks act as delegated monitors of borrowers under both free-entry and monopoly regimes.

The delegated monitoring model, introduced by Diamond (1984), assumes that investors can only observe ("monitor") a borrower's returns or actions at some cost. There is incentive to delegate the monitoring function to an intermediary ("bank"), but the bank must also be monitored by its depositors. The main result is that a sufficiently large and diversified bank dominates direct lending because its deposits have low default risk and require little monitoring.

Thus, like other models of bank formation, delegated monitoring implies that bank performance improves with diversification.[2] Although this suggests finite banks have economies of scale, papers that address banking in general equilibrium (e.g., Boyd and Prescott (1986) and Williamson (1986)) assume that the economy is infinite and allows complete diversification of bank portfolio risk, so that an infinite number of riskless banks form and compete. Since actual banks are finite and risky, I model equilibrium behavior of delegated monitors in a finite, risky environment.

My model modifies that of Williamson (1986) by making the total number of agents in the economy finite and by determining the relative numbers of direct investors and borrowers endogenously. Sections 2 and 3 describe the economy and its equilibria under a direct lending regime.

Section 4 examines banks as a mechanism for delegated monitoring. Banks can dominate direct lending either by diversifying or by using their own capital to reduce their default risk. Increases in bank size improve diversification but weaken capitalization, leading to ambiguous relationships between bank size and agents' returns. Also, because diversification reduces expected bank profits per depositor, a bank may not wish to expand unless it can lower its deposit interest rate.

Section 5 examines equilibria when investors can freely choose to become banks. At least one such "free-entry" intermediated equilibrium exists whenever direct lending is feasible, and free-entry equilibria also exist in situations where direct lending is less attractive than immediate consumption ("autarky"). On the other hand, there are generally multiple free-entry equilibria. Although some of these equilibria vary by bank size, they may not be Pareto-ranked by bank size because of the conflicting forces of diversification and capitalization already mentioned.

Multiple equilibria are caused by an adoption externality: a depositor's return depends not only on the terms offered by her bank but on the *size*

[1] Bhattacharya and Thakor (1993) survey both areas of research.

[2] For references on diversification in banking, see Winton (1994).

of her bank, which in turn depends on how many other agents choose to use that bank. Although some free-entry equilibria dominate others, achieving a Pareto-optimal equilibrium requires coordination by all parties: banks, depositors, and firms.

A possible hindrance to this is a form of Myers' (1977) debt overhang problem: in several situations, banks may not wish to deviate from a dominated equilibrium without lowering the nominal rate on depositors' debt contracts, because otherwise the depositors get too much of the gain from the change. For example, some intermediated equilibria actually increase credit rationing compared to the economy's direct-lending equilibrium.[3] Although a unilateral increase in a bank's loan rate may reduce rationing, this may not be profitable for the bank unless it can lower its deposit rate. Similarly, even if the benefits of diversification cause equilibria with small banks to be dominated by an equilibrium with larger banks, small banks may not wish to increase their size unless they can do it by lowering their deposit rate.

Such problems cast doubt on whether the economy can easily achieve an optimal banking structure. Of course, this paper's single-period noncooperative framework leaves no room for the historical evolution of beliefs or communication and cooperation among agents.

Concerns about market imperfections in the banking sector have often resulted in regulatory intervention, a common form of which has been entry restrictions. In this model, if larger banks dominate smaller ones, entry restrictions might have the benefit of focusing agents on a few banks; the drawback is that such restrictions may make it easier for banks to collude.

Section 6 examines the situation where the government allows only one bank to operate. The monopoly bank tries to exploit its position subject to the constraint that investors and firms can bypass the bank and contract directly with one another. Because firms bear a sunk cost in acquiring the production technology, the bank may exploit firms to the point where rational agents prefer to forego entrepreneurship. Thus, a monopoly bank may be unable to overcome autarky even when intermediation with free entry is feasible. In addition, a monopoly bank in a large economy may exploit firms so much that their number is reduced relative to the direct lending regime; in fact, such a bank can only increase equilibrium production in a subset of the economies in which direct lending leads to credit rationing.

Other authors also address the issue of competition among finite delegated monitors. Krasa and Kubitschek (1988) focus on the role of differ-

[3] Here, "credit rationing" is used in the sense of Williamson (1986): it occurs when the highest return investors can obtain under an incentive compatible loan contract leaves the supply of funds strictly lower than demand.

ential monitoring costs in determining financial sector structure; in their model, multiple equilibria are possible, in some of which direct lending by low-cost investors coexists with intermediation for high-cost investors. Daltung (1991) assumes that it is firms rather than investors which act as intermediaries; since a firm acting as bank need not monitor itself, there may be welfare gains to having more than one bank.[4] Yanelle (1988) shows that, when loan and deposit markets take place sequentially, the precise ordering of these markets has a dramatic effect on competitive outcomes.

These papers share several features that differ from this one. All fix the numbers of firms and investors such that the supply of funds exceeds demand; all have lending and borrowing markets occurring sequentially rather than simultaneously, and all except for Daltung have agents within a given market acting sequentially as well; finally, all ignore bank capital. Although they permit strong conclusions, these assumptions seem questionable. Oversupply of funds and lack of bank capital are not equilibrium outcomes if agents can choose their own roles, and it is difficult to see why either the lending or the borrowing market should necessarily occur before the other. For these reasons, my paper provides a richer examination of equilibrium banking in a finite economy.

These differences in modeling lead to very different results. The papers just described often find that, in equilibrium, the banking sector is a monopoly. By contrast, because my paper emphasizes the role of bank capital as an alternative to diversification, it predicts that, absent regulation to the contrary, the banking sector may have several banks. This prediction should be strongest in a setting where capital is an important factor and additional ("outside") capital is expensive (see Smith (1986) for a survey of evidence on the costs of issuing equity). Thus the higher are information costs, the bigger the gap between "inside" and outside capital, and the more likely that a number of banks coexist. The stylized facts of "free banking" periods in various countries are consistent with this prediction: such periods were characterized by a substantial number of banks in equilibrium, and, in all cases, these periods occurred well before technological changes in the late 20th century substantially reduced the costs of information (see Schuler (1992)).

A further prediction concerns banks' ability to diversify risk. If loan returns have an unobservable systematic component, gains to diversification are reduced. Thus, economies with more systematic risk should see greater emphasis on bank capital and less on size.[5]

[4] Despite this potential welfare gain from multiple banks, Daltung assumes that investors choose larger banks over smaller ones, resulting in a natural monopoly in banking.

[5] Krasa and Villamil (1992b) explore a similar tradeoff in a model where banks have no capital, but the cost of monitoring a failed bank increases with the bank's size. As the relative size of the systematic risk component increases, optimal bank size decreases.

2. BASIC ASSUMPTIONS

M agents live in an economy lasting two periods, 0 and 1. All agents have utility $y_0 - e_0 + (y_1 - e_1)/\beta$, where y_t is period-t consumption, e_t is period-t effort, and $\beta \geq 1$ is the common discount rate; agents have an unlimited supply of effort and begin life with one unit of a nonstorable consumption good.

At the beginning of period 0, any agent can acquire a fixed-scale production technology at the cost of one unit of the consumption good; the acquisition is publicly observable.[6] The technology converts K units of period-0 consumption into $K\tilde{X}$ units of period-1 consumption; here, K is an integer greater than one, and $\tilde{X}$ is a random variable with positive support $[0, X_m]$, continuous density $f(x)$, and distribution $F(x)$. Returns are stochastically independent across any agents using the technology.

Any agent observes her own return costlessly, but others must expend c units of effort in order to observe (monitor) the return, and such observations are private information. Since any agent ("firm") who acquires the technology must obtain K units of consumption from other agents ("investors") in period 0, contracts between investors and firms must address the chance that the firm may misrepresent its returns.

Assuming for now that only direct contracting between firms and investors is possible, the order of events in the economy is as follows:

1. Agents decide whether or not to become firms (acquire the technology).
2. Firms contract with investors for funds. Contracts specify amounts invested, firm announcements that cause investors to monitor, and payments to investors based on announcements and the results of any monitoring.
3. Investment and/or consumption take place. Period 0 ends.
4. Firm returns are realized.
5. Firms announce results.
6. Monitoring takes place as specified by contract.
7. Proceeds are divided according to contract and consumed. Period 1 ends.

3. EQUILIBRIUM WITHOUT INTERMEDIATION

Determining equilibrium in this direct-lending economy requires three elements: the contracts used between firms and investors, how firms and

[6] The assumption that acquiring the technology uses up one's goods endowment simplifies the analysis, but all the model really requires is that acquiring the technology involves some sunk cost in goods or effort.

investors come to agree on specific terms, and how agents choose whether to become firms or investors. I begin with the choice of contracts.

As noted, a firm can misrepresent its returns to investors. Although investors can detect misrepresentation by monitoring, this is costly; optimal contracts will be designed to minimize the costs of such monitoring. Since this contracting problem has been extensively studied by Townsend (1979) and others, I make use of their results.

For simplicity, assume that investors can commit to monitor as a deterministic function of a firm's announced return, and that a firm uses the same contract with all of its investors. Then the optimal incentive compatible contract is what Gale and Hellwig (1985) call a "standard debt contract," in which each investor receives a fixed amount ("stated rate") R if the realized return x is greater than or equal to R, and verifies and receives x otherwise.[7] Since any other symmetric contracts would leave gains to renegotiation between investors and firms, I assume that only standard debt contracts are used.

A firm that borrows at stated rate R has expected return

$$\pi_f(R) \equiv K \cdot \int_R^{X_m} (x - R) f(x)\, dx, \tag{3.1}$$

while its investors each have expected return

$$r(R) \equiv \int_0^R x f(x)\, dx + R[1 - F(R)] - cF(R). \tag{3.2}$$

Note that $\pi_f(R)$ is strictly decreasing in R. For simplicity, assume $r(R)$ is strictly quasi-concave in R, so that it has a unique global and local maximum.[8] Call the corresponding argmax R^*. It is possible that R^* is less than X_m, so that $\pi_f(R^*) > 0$. The higher the monitoring cost c, the greater the chance that R^* is such an interior point, ceteris paribus.

[7] Townsend (1979) shows that incentive compatibility requires a fixed payment to an investor for unverified announcements, and lower payments for any verified announcements. Specializing to risk neutral agents, Gale and Hellwig (1985) show that, because "standard debt" compresses monitoring into the smallest possible region, it dominates all incentive compatible contracts. Williamson (1986) extends these arguments to the case of multiple investors and symmetric contracting. The optimal contract changes if the assumptions that monitoring is deterministic and that all investors receive the same contract are relaxed; see Mookherjee and Png (1989) and Winton (1995).

[8] It can be shown $r(R)$ is quasi-concave if $F(\cdot)$ has monotone increasing hazard rate $f(\cdot)/[1 - F(\cdot)]$. Examples of distributions with this property include the uniform, exponential, normal, gamma with parameter greater than one, and their truncated versions. See Barlow and Proschan (1975).

Let μ_R denote the expected payment ("cash flow") to an investor under a debt contract with stated rate R. This can be written as

$$\mu_R \equiv \int_0^R xf(x)\,dx + R[1 - F(R)] = R - \int_0^R F(x)\,dx, \qquad (3.3)$$

where the second equality follows from integration by parts. Note that the integral term in the third expression equals expected shortfalls from the stated rate on the debt.

Since the stated rate R summarizes a loan contract's terms, the next question is how R is determined. Since this should depend on the relative bargaining power of firms and investors, I make the following assumptions. If there are n firms and more than nK investors with funds, competitive pressure bids R down to the lowest rate R^o such that investors receive the same discounted return they would get from consuming their own endowment—that is, R^o is the lowest R such that $r(R) = \beta$ (assuming R^o exists). Conversely, if nK investors have funds and there are more than n firms, competitive pressure bids R up to R^*, maximizing investor returns, and firms are randomly rationed. If there are exactly enough investors to fund all firms, the rate R can be anywhere between R^o and R^*, depending on agents' beliefs; although there is no reason rates could not differ across firms, I focus on symmetric outcomes.[9]

The last step in determining equilibrium is analyzing how agents choose whether or not to acquire the technology. Since relative numbers of firms and investors determine the loan rate R, which in turn determines relative returns, any agent's choice depends on her beliefs about other agents' choices. To keep matters simple, I focus on equilibria in which agents do not randomize in their decision to acquire the technology. An equilibrium with N firms and $M - N$ investors assumes that N agents (perhaps those indexed by $1, \ldots, N$) choose to acquire the technology, the others do not, and each agent knows this.[10]

[9] Similar results could hold if there are more than n firms and more than nK and less than $(n + 1)K$ investors, but this case is ruled out below for simplicity.

Although the outcomes I have chosen seem most plausible, the fixed scale of production permits all manner of equilibria. For example, if there are $n + 1$ firms and nK investors, a rate of R^o could be sustained in equilibrium: even if the unfunded firm offered investors a higher rate, no single investor would take the offer unless she believed $K - 1$ others were also willing to do so.

[10] Although randomized choice might be more plausible in a large economy, this would cause stochastic imbalances between investors and firms, making exact calculation of equilibrium difficult. Since the impact of these imbalances would be relatively small for $M/(K + 1)$ at all large, I focus on deterministic choices. Similar considerations apply for intermediated equilibrium (see Section 5).

Accordingly, define a *direct-lending equilibrium* (DE) as a number of firms N^d, a number of funded firms n^d, and a stated rate R^d such that (i) in the second stage of the economy, R^d is determined by relative numbers of firms and investors as described above, and (ii) in the first stage of the economy, given the equilibrium actions of other agents, no agent prefers to change her choice of whether or not to acquire the technology.

The following proposition spells out the three basic types of direct lending equilibria: autarkic, rationed, and nonrationed.

PROPOSITION 3.1. (i) *If* (a) $r(R^*) < \beta$, *or* (b) $\pi_f(R^o) < \beta$, *then the only direct-lending equilibrium of the economy is* autarkic: $N^d = n^d = 0$, *and* R^d *is undefined (no one acquires the technology and all consume their endowment).* (ii) *If* $\beta \le r(R^*) < \pi_f(R^*)$, *then the equilibrium is* rationed: $n^d < N^d = (M - n^d)/K$, $R^d = R^*$, *and* $(n^d/N^d) \cdot \pi_f(R^*) \approx r(R^*)$. (iii) *If* $\beta \le \pi_f(R^o) \le \pi_f(R^*) \le r(R^*)$, *the equilibrium is* nonrationed: $n^d = N^d = (M - n^d)/K$, $R^d \le R^*$, *and* $\pi_f(R^d) \approx r(R^d)$.[11]

Proofs of all results may be found in the Appendix unless otherwise stated. The intuition for Proposition 3.1 is straightforward. In part (i), if (a) holds, the best contract firms can offer leaves investors worse off than autarky; if (b) holds, investors cannot be given a better than autarkic return without giving firms a worse than autarkic return. If the condition in (ii) holds, any funded firm must be strictly better off than any investor; investors will choose to become firms until the expected return to a firm (including rationing) equals that to an investor. Finally, condition (iii) assures that it is possible to find a loan rate at which funded firms and investors earn the same return, so that no one has incentive to switch roles in the first stage of the economy.[12]

4. INTERMEDIATION

In the direct-lending economy, every funded firm is monitored by K investors, each of whom incurs expected costs of $cF(R)$. This duplication of costs creates an incentive to have one agent monitor on behalf of the

[11] Here and in the rest of the paper, I assume that K evenly divides $M - N^d$, so as to avoid integer effects where one or two extra agents have a large effect on equilibrium rates. I also ignore the case where $r(R^*) < \pi_f(R^*)$ but M is so small that no investor finds it worthwhile to deviate and become a firm.

[12] Generically, nonautarkic equilibria do not maximize total welfare subject to the incentive compatibility constraints. The direct lending allocation that maximizes welfare has $n^d = N^d = M/(K + 1)$ firms and a rate of R^o, maximizing production and minimizing monitoring costs. Although this allocation gives firms higher returns than investors, it would Pareto-dominate the direct-lending equilibrium on an ex ante basis if agents' roles could be assigned randomly by a central planner.

investors; however, the agent could mislead the investors, and the cost of an incentive compatible contract might outweigh the direct savings from delegation.

One possibility is to have an investor act as intermediary ("bank"), borrowing funds from other investors ("depositors") and lending these funds and its own endowment to firms. Having invested on its own account, the bank has bound itself to monitor the firms, and the contract between the bank and the depositors provides incentives for it to tell them the truth. For simplicity, assume that loans and deposits are both standard debt contracts.[13]

The result derived in Diamond (1984) and Williamson (1986) is that a large enough bank can dominate direct lending. More formally,

PROPOSITION 4.1. *Take any rate R such that direct lending at R weakly Pareto-dominates autarky. Then there exists an n' such that, for all $n > n'$, a bank that lends to n firms and borrows from $nK - 1$ depositors can offer contracts that strictly Pareto-dominate direct lending at R.*

For a formal proof in this setting, see Williamson (1986). The intuition is that, as the number of depositors goes to infinity, the Strong Law of Large Numbers implies that the bank's portfolio return per depositor converges to its mean almost surely. As long as the stated rate on the bank's deposits is less than this mean return, the expected return to depositors converges to the stated deposit rate. Consequently, a sufficiently large bank can set a deposit rate that assures both the bank and its depositors of expected returns greater than those obtained from direct lending at the bank's loan rate. Since at any given rate firms are indifferent between direct loans and bank loans, a small decrease in the loan rate makes everyone better off.

Nevertheless, diversification across large numbers of projects is not the only way to make delegated monitoring feasible:

PROPOSITION 4.2. *Take any rate R such that direct lending at R weakly Pareto-dominates autarky. Then a bank that borrows from $K - 1$*

[13] Because banks compete simultaneously for loans and deposits, and depositors have rational expectations about bank size, depositors know the bank's portfolio return distribution. Thus the results cited in the previous section suggest that the optimal symmetric deposit contract is still standard debt. (If depositors did not observe the form of the bank's loan contracts, they might prefer nondebt deposit contracts so as to forestall risk-shifting by the bank.)

Optimal loan contracts are harder to pin down. A nondebt loan contract increases bank portfolio risk in the sense of Rothschild and Stiglitz (1970). If deposits are debt contracts, such risk-shifting increases the bank's share of loan payments at its depositors' expense, but it also increases the bank's monitoring costs. Krasa and Villamil (1992a) show that a sufficiently large bank prefers debt for its loan contracts; Winton (1990) shows that the same is true for a bank with a single loan.

depositors and lends to one firm can offer contracts that strictly Pareto-dominate direct lending at R.

The intuition is that the bank's own endowment ("capital") provides a cushion for depositors, preventing bank default and thus depositor monitoring in some states where the firm cannot make good on its bank loan. A bank could raise more capital by issuing subordinated contracts, but investors in these contracts would have to monitor, making this capital more expensive than the bank's own endowment.[14] For simplicity, I limit bank capital to the latter amount.

Thus, capital and diversification are two mechanisms for implementing delegated monitoring. Since the bank's capital is fixed, increases in bank size strengthen diversification but weaken relative capitalization. Because firm returns are assumed to be pairwise independent, sufficient diversification can drive the default probability to zero. Nevertheless, smaller banks could dominate somewhat larger ones, and this would be more likely to hold if firm returns had an unobservable systematic component.

Since a bank uses debt contracts for both deposits and loans, its structure is defined by three variables: deposit rate, loan rate, and size. I now examine how these variables affect the various agents' objective functions.

Suppose a bank borrows from $nK - 1$ depositors at rate D and invests in n firms at rate R. Let $\pi_b(R, D; n)$, $\pi_f(R, D; n)$, and $r_b(R, D; n)$ be the corresponding expected returns to the bank, each firm, and each depositor, respectively. Since a loan to firm i at rate R returns $K \cdot \min\{\tilde{X}_i, R\}$, the total return to the bank's portfolio is $K \cdot \Sigma_i (\min\{\tilde{X}_i, R\})$. Define this return as $\tilde{W}_n(R)$, and let $\tilde{Y}_n(R) = \tilde{W}_n(R)/(nK - 1)$ denote the bank's portfolio return per depositor. Let $h_n(w; R)$ and $H_n(w; R)$ denote the density and distribution functions of $\tilde{W}_n(R)$, and let $g_n(y; R)$ and $G_n(y; R)$ denote those of $\tilde{Y}_n(R)$. Note that $G_n(y; R) = H_n[y(nK - 1); R]$.

For any stated loan rate, a firm's objective function is the same whether it writes K debt contracts with individual investors or one debt contract with a bank. Thus $\pi_f(R, D; n) \equiv \pi_f(R)$ as given in (3.1), and firm returns are not directly affected by changes in D or n. Depositors get the expectation of the smaller of D or $\tilde{Y}_n(R)$ less the expected cost of monitoring the bank, and the bank gets its expected loan portfolio return less expected payments to depositors and costs of monitoring firms. Using the expres-

[14] Because the bank's equity claim on its single loan and the depositors' debt claims are, respectively, isomorphic to junior and senior debt issued directly by the firm, Proposition 4.2 is analogous to the result in Winton (1995), where debt contracts of varying seniority dominate a single class of debt. By extension, if the bank raised capital from other investors, the optimal contracts would resemble junior debt or preferred stock rather than common stock.

sion for the expected payment received by a debtholder, it follows that

$$r_b(R, D; n) = D - \int_0^D G_n(y; R)\, dy - cG_n(D; R), \text{ and} \tag{4.1}$$

$$\pi_b(R, D; n) = nK\mu_R - (nK - 1)D + \int_0^{(nK-1)D} H_n(w; R)\, dw - ncF(R), \tag{4.2}$$

where again μ_R is the expected payment on a one-unit loan to a firm at rate R.

Lemma A.1 in the Appendix establishes several properties of the bank's portfolio distribution. Of these, the most important is that, for a bank of any given size, an increase in the bank's loan rate R causes a first-order stochastic dominance shift in its portfolio distribution $H_n(\cdot)$. This helps establish the following lemmas on the impact of changing bank lending and deposit rates.

LEMMA 4.3. *Ceteris paribus, increasing the loan rate R* (i) *decreases firm profits $\pi_f(R)$,* (ii) *increases depositor returns $r_b(R, D; n)$, and* (iii) *increases both the bank's expected cash flow and its cost of monitoring firms, with an ambiguous net effect on the bank's total return $\pi_b(R, D; n)$.*

LEMMA 4.4. *Ceteris paribus, increasing the bank's deposit rate D* (i) *increases the depositors' expected cash flow and increases their cost of monitoring the bank, with an ambiguous net effect on $r_b(R, D; n)$, and* (ii) *decreases the bank's expected profits $\pi_b(R, D; n)$. In addition,* (iii) *$r_b(R, D; n)$ unambiguously increases with D if either* (a) *n equals 1 and $D \leq$* $\min\{KR^*/(K-1), KR/(K-1)\}$, *or* (b) *$n$ is sufficiently large and D is less than μ_R.*

The effects of a change in bank size n on the returns to a bank and its depositors are more ambiguous. For depositors, this is due in part to the interplay between diversification and leverage as n increases; depending on which is stronger, the expected payment received by depositors increases or decreases. Increases in bank size can also have an ambiguous effect on depositors' costs of monitoring: even ignoring changes in bank capitalization, it is easy to construct examples where an increase in bank size increases a bank's chance of default.[15]

Similar considerations lead to ambiguous effects of size on bank profits. To the extent diversification or leverage predominates, the bank's per depositor profit margin tends to decrease or increase; however, the change in size also has a direct effect on bank profits. The following

[15] Winton (1990) provides a more detailed analysis of bank size effects on depositor cash flows and monitoring costs.

lemma describes one situation in which bank profits actually decrease with size:

LEMMA 4.5. *Suppose R, D, and n are such that $\pi_b(R, D; n) = r_b(R, D; n)$. If an increase in bank size to n' would cause the expected payment to a depositor to increase, or to decrease by a sufficiently small amount, then $\pi_b(R, D; n')$ is smaller than $\pi_b(R, D; n)$.*

In other words, if expected returns to banks and depositors are equal, and an increase in size improves expected depositor cash flows, a bank will not willingly increase its size unless it can lower its deposit rate. The intuition is that if greater size implies larger expected payments to depositors, the bank's net profits per depositor fall. Because the bank increases its leverage (it cannot raise more capital), payments to depositors are a greater fraction of bank lending revenues as well. These two effects and the assumption that the bank's initial profits equal a depositor's initial expected return are enough to imply that the bank's profits decrease. As the next section shows, this can lead to cases in which banks refuse to increase their size even though larger banks are potentially more efficient.

5. EQUILIBRIUM WITH FREE ENTRY INTO BANKING

Having examined the mechanism of intermediation in a finite setting, I now turn to the equilibrium behavior of agents when intermediation is allowed. This requires a description of the order of events in this setting:

1. Agents decide whether or not to become firms.
2. Investors that wish to act as banks offer contracts to other investors and firms.
3. Investors and firms choose whether to use a bank or contract with one another directly.
4. Investment and/or consumption take place. Period 0 ends.
5. Firm returns are realized.
6. Firms announce results.
7. Monitoring takes place as specified by contract.
8. Proceeds are divided according to contract and consumed. Period 1 ends.

As in the direct-lending economy, equilibrium consists of a subgame perfect Nash equilibrium of the game where agents first choose whether or not to become firms, then decide what contracts to offer (given their choices), and finally decide which contracts to accept. Matters are more complicated now, since those investors who wish to act as banks make offers to both sides of the market; however, equilibrium assumes that each agent knows other agents' strategy choices and makes her own decisions accordingly.[16]

Agents' beliefs about other agents' actions are critical. Because of the fixed project scale, deviation by any agent results in autarky for that agent unless others join her; thus, a large number of Nash equilibria exist, including direct lending and autarky. It is also possible that one bank captures the market despite rival banks who post deposit rates that are *potentially* more attractive: the returns depositors earn at the rival banks may not be higher if these banks only attract a small number of investors and firms. In general, adoption externalities and multiple equilibria are an inescapable part of this model.[17]

Since agents are free to choose roles at the outset, and investors can always freely offer to act as banks, I focus on equilibria in which all agents earn the same expected returns. Since (ignoring discreteness) an outcome which failed to meet this criterion would give agents incentive to switch roles, this should be a necessary condition for equilibrium in an economy where individual agents are relatively "small." In addition, as in direct-lending equilibrium, I require that individual firms cannot make offers with the potential to make agents better off. I also make several simplifying assumptions. First, as already noted, there is no interbank borrowing or lending. Second, banks only compete through their stated rates—they cannot voluntarily limit their size, although passive rationing due to imbalances may occur. Finally, as in the direct-lending economy, I focus on equilibria in which agents choose their roles (firm, bank, or depositor) deterministically.[18]

Formally, a *free-entry intermediated equilibrium* (FE) is defined as a number of firms N^e, bank size n^e, number of active banks B^e, and a pair of rates R^e and D^e such that: (i) $M - N^e = n^eK \cdot B^e$; (ii) when banks lend to n^e firms at R^e and borrow from $n^eK - 1$ depositors at D^e, bank and depositor expected returns equal each other and exceed the autarkic return β; (iii) expected returns are such that, fixing all other agents' choices, no investor wishes to become a firm, or vice versa; (iv) no firm can offer K

[16] The order of events does not allow interbank contracting, ruling out equilibria where firms choose one set of banks, depositors choose another, and the first set of banks borrows from the second. These issues are left for future research.

[17] Yanelle (1988) reduces the number of equilibria by having either lending or borrowing activity take place before the other. Such assumptions seem unnatural and produce results that vary greatly with the precise order of borrowing and lending activity.

[18] Randomized choice of roles by agents would cause stochastic imbalances, which could be met by rationing or interbank lending. In either case, bank size would be random; because the effect of bank size on bank and depositor returns is complex, the analysis of intermediated equilibria would be intractable.

investors a direct loan rate R such that both the firm and the investors prefer it to their intermediated returns.

A free-entry intermediated equilibrium is in fact subgame perfect: because autarky is always self-fulfilling, simply assume that out of equilibrium actions result in autarky. Several properties of such equilibria can now be derived:

PROPOSITION 5.1. *If an economy has a nonautarkic direct-lending equilibrium, it always has at least one free-entry intermediated equilibrium strictly dominating direct lending.*

This follows directly from Proposition 4.2. A bank with only one loan can write contracts strictly dominating direct lending at any rate R, setting its deposit rate so as to bring its expected returns into line with those of its depositors. It is then simply a matter of setting its loan rate so that expected returns to firms equal those to the bank and to the depositors.

PROPOSITION 5.2. *A sufficiently large economy that has a nonautarkic direct-lending equilibrium has multiple free-entry intermediated equilibria.*

This is a direct result of Propositions 4.1 and 5.1. A sufficiently large bank can dominate direct lending, and so, as the economy's size increases, there are FEs with one or more large banks. However, comparing aggregate welfare across different FEs is difficult because of the offsetting leverage and diversification effects described in Section 4. An equilibrium that has relatively small banks may well be more efficient than one with somewhat larger banks, particularly if noncontractible systematic risk is present.

The next result concerns the existence of a free-entry intermediated equilibrium when direct lending leads to autarky:

COROLLARY 5.3. *Suppose the economy's direct-lending equilibrium results in autarky, so that either $\pi_f(R_0) < \beta$ or $r(R^*) < \beta$.* (i) *For any size economy, there are cases when either $\pi_f(R_0) < \beta$ or $r(R^*) < \beta$ and a free-entry intermediated equilibrium exists.* (ii) *Define R^{**} as the argmax of $\mu_R - (c/K)F(R)$. If the economy is sufficiently large, free-entry intermediated equilibria exist if there is an $R' < R^{**}$ such that $\mu_{R'} - (C/K) \cdot F(R') = \beta$ and $\pi_f(R') > \beta$.*

The intuition for part (i) follows from Proposition 5.1: an FE where banks have one loan each always exists if direct lending is feasible. Continuity arguments imply that there are economies in which direct lending is just infeasible yet the FE exists. Part (ii) follows because a sufficiently large bank with loan rate R' and deposit rate β offers depositors a return arbitrarily close to β, while the bank's own return also converges to β.

Since $\pi_f(R')$ exceeds β, this arrangement dominates autarky, and lending and deposit rates can be increased until all agents' returns are equal.

Although intermediation with free entry overcomes autarky more often than does direct lending, free entry may not achieve banks of the most efficient size. Also, even though the definition of an FE implies it Pareto-dominates the direct-lending equilibrium, FEs may actually introduce or increase credit rationing:

PROPOSITION 5.4. *Suppose the direct-lending equilibrium of the economy involves loan rate R^d, number of firms N^d, and number of funded firms n^d. Then* (i) *if $R^e > R^d$, an FE must reduce credit rationing* ($B^e \cdot n^e > n^d$ *and* $N^e < N^d$); (ii) *if $R^e \leq R^d$, an FE can increase credit rationing.*

When R^e exceeds R^d, profits of funded firms are lower under intermediation than under direct lending. Since both equilibria equate the expected returns of all types of agents, and the FE dominates direct lending, it is easy to show that rationing must be lower with intermediation. On the other hand, because it reduces monitoring costs, intermediation may improve investors' returns even when the loan rate R^e is lower than the direct loan rate R^d; since this implies funded firms are better off than under direct lending, it is possible that credit rationing increases relative to direct lending.

To summarize, multiple FEs exist, some vary by bank size, and some may increase credit rationing compared to direct lending. This variety of outcomes occurs because an FE only requires that all agents earn the same expected returns and that firms cannot make dominating offers; once more, *any* configuration that satisfies these requirements can be supported as an FE by assuming that deviations result in an autarkic outcome. Since this "off-the-path" behavior is extreme, plausible restrictions on out-of-equilibrium behavior could significantly reduce the number of possible FE outcomes.

One might argue that, just as an FE requires that firms cannot make offers that entice depositors away from banks, it should also require that banks cannot make offers that Pareto-improve upon the equilibrium outcomes. The difficulty lies in deciding how agents would react to such a deviation. When a firm deviates, all that is required for success is that K investors respond; furthermore, all know exactly what returns will be if K investors do come. By contrast, the bank's offer requires $nK - 1$ investors for every n firms, and may not dominate the equilibrium in question if n is not within some range. Thus, coordination difficulties are likely to be larger for banks than for firms.

It is difficult to resolve this issue in the single-period noncooperative setting of this model, since by definition this excludes any historical basis for expectations and any communication and coordination among agents.

Nevertheless, some further insight can be gained by focusing on an individual bank's incentives to deviate from certain dominated free-entry intermediated equilibria. I examine two cases: FEs in which credit rationing is higher than in the economy's direct-lending equilibrium, and FEs in which bank size is smaller than optimal.

The next proposition shows that, if an FE increases credit rationing over direct lending, Pareto-improvement is always feasible, but this may require actions that no bank can undertake unilaterally.

PROPOSITION 5.5. *Suppose an FE involves more rationing of firms than the direct-lending equilibrium of the economy:* $B^e \cdot n^e < n^d$ *and* $N^e > N^d$. (i) *Then any bank can offer its depositors a rate* $D \leq D^e$ *and offer* n^e *firms guaranteed funding at a rate* $R > R^e$ *such that, if these offers are accepted, the bank, its depositors, and the firms it funds are all strictly better off than under the FE.* (ii) *The bank can improve welfare in this fashion without changing its deposit rate from* D^e *if either* (a) *bank size* n^e *is one, or* (b) n^e *is sufficiently large and* $D^e < \mu_{R^e}$.

The intuition for (i) is that, if intermediation increases rationing, then $R^e < R^d$, and a slight increase in a bank's loan rate improves depositor returns and total bank and depositor welfare, ceteris paribus. Since the firms go from being rationed to getting loans for certain at a slightly higher interest rate, they are better off as well. Still, the increased cost of monitoring caused by the higher loan rate may exceed the fraction of the increase in cash flows which the bank keeps for itself. This is an example of the debt overhang problem first analyzed by Myers (1977): while the total net value of the higher loan rate is positive, debtholders (depositors) obtain part of the benefit. If this is the case, a suitable decrease in the deposit rate would leave both the bank and depositors better off than in the original FE.

Part (ii) of the proposition shows that debt overhang is not a problem for very small or sufficiently "well-diversified" banks. On the other hand, a bank of intermediate size may not wish to deviate from the FE unless it can be assured that depositors will accept a lower deposit rate D.

The other "problem" I consider is one in which the economy has (at least) two FEs, one of which has larger banks than, and Pareto-dominates, the other. Assume for simplicity that neither equilibrium involves credit rationing. The question is, if agent beliefs initially favor the FE with smaller banks, does any bank have incentive to deviate and try to increase its size?

The answer to this question is likely to be no if the bank tries to increase its size by offering the same or higher deposit rate and lowering its loan rate. In the base FE, banks, firms and depositors earn the same expected returns. If a larger bank is to Pareto-dominate this outcome, it must offer firms a loan rate that is no higher than the one they enjoy in equilibrium. In addition, if depositors are better off with a larger bank, it is likely that diversification benefits outweigh decreases in relative capitalization. This is precisely the situation where Lemma 4.5 suggests that the bank will not be better off unless it lowers its deposit rate as well. Once again, this follows from a kind of debt overhang: at the equilibrium deposit rate, increases in diversification help depositor cash flows and hurt the bank on a "per dollar" basis, so the bank only wants to increase size if it can lower its rate—an action which seems unlikely to attract more depositors.

The point of both of these examples is that socially desirable bank actions on a single side of the market (lending) may not be in the bank's interest absent rate "concessions" from depositors. These issues have not been noted in earlier work because this focused on perfectly diversified banks, where increases in size no longer help performance and, as Proposition 5.5 shows, debt overhang does not prevent banks from raising loan rates to reduce rationing. Because finite banks have risky deposits, lending decisions affect depositors, and the resulting externality enhances the need for coordination among agents.

Once again, it is important to recognize the limitations of this approach. Coordination difficulties may be resolved to some extent by multiperiod play and communication among agents; if an economy has banks which later prove to be "too small," mergers may resolve this in a way that induces depositors to accept a lower rate. Still, the results of this section show that an intermediated economy does have the potential for multiple equilibria, and that unilateral action by banks may not overcome the coordination problems that result.

6. EQUILIBRIUM WITH MONOPOLY BANKING

If coordination problems are a concern, and larger banks Pareto-dominate smaller ones, regulators might be tempted to limit entry in an effort to focus activity on a few institutions. A number of governments have at various times imposed such restrictions on their banks, although their reasons have often focused more on the money creation role of banks rather than their lending role. Of course, entry restrictions may support collusive behavior by banks.

To explore the impact of such restrictions on an intermediated economy, this section examines the extreme case where only one bank is allowed to operate. I assume that regulators do not restrict bank lending or deposit rates, and that direct lending between firms and investors is still possible. However, to simplify, I assume that the bank is not allowed to ration loans or deposits unless there is an oversupply of one or the other, and that the rationing rule is random.[19] Thus, if there are N firms in the economy, then the bank makes n loans and has $nK - 1$ depositors, where n equals the smaller of N and $(M - N)/K$. I assume that depositors believe everyone will use the bank if such an outcome gives higher returns than does direct lending or autarky. Such "coordinated" beliefs seem plausible in a setting where everyone knows there is only one bank in the economy.

Assume once more that a firm that offers potentially better terms than other borrowers does in fact attract exactly K investors. Then a viable bank must offer depositors a return at least as great as the larger of the investors' autarkic return β or the expected return earned by lending directly to a firm. If a firm can obtain (possibly rationed) loans from the bank at rate R, it will not borrow directly at any rate above $R^{f}(R)$, where $R^{f}(R)$ is the unique solution to $\pi_f(R^f) = \pi_f(R) \cdot (n/N)$. This means that, if it is to obtain funds, the bank must ensure that direct lending at rates below $R^f(R)$ leaves investors worse off than going to the bank and receiving a (possibly rationed) deposit contract with stated rate D. From Section 3, a depositor cannot earn more than $\mu_{R^*} - cF(R^*) \equiv r(R^*)$ from direct lending, even if $R^f(R)$ exceeds R^*. Thus, the bank must offer depositors a return of at least $r_b^m(\mathrm{R}) \equiv \max\{\beta, r(\min\{R^f(R), R^*\})\}$.[20]

The bank maximizes its profits subject to the depositors' return constraint just described. Since bank profits decrease in the deposit rate D, given any loan rate R, the bank will choose the lowest D such that the constraint holds exactly; call this rate $D^m(R, n) \equiv D^m$. Thus, the bank's problem becomes

$$\max_R \; nK\mu_R - (nK - 1) \cdot \left[D^m - \int_0^{D^m} G_n(y; R)\, dy\right] - ncF(R), \quad (6.1)$$

where D^m is the smallest solution D to

$$\delta \cdot \left[D - \int_0^D G_n(y; R)\, dy - cG_n(D; R)\right] + (1 - \delta) \cdot \beta = r_b^m(R) \quad (6.2)$$

and $\delta \equiv (nK - 1)/(M - N - 1)$ is the chance that a depositor will *not* be rationed. Substituting from (6.2), the bank's objective function can be

[19] Ex post, rationed firms are willing to borrow at any rate, and rationed depositors are willing to lend directly at any rate above R^o. Binding random rationing rules prevent a few extra firms or depositors from causing vast swings in equilibrium deposit and loan rates. The end of this section discusses the bank's incentive to arbitrarily limit its size.

[20] If firms had to ration investors, the bank would have to offer a return of at least a convex combination of $r_b^m(R)$ and β, but the analysis in this section would be essentially unchanged.

restated as

$$nK \cdot [\mu_R - (c/K)F(R)] - (nK - 1) \cdot \{\delta^{-1} \cdot [r_b^m(R) - (1 - \delta)\beta] + cG_n(D^m; R)\}. \tag{6.3}$$

The following lemma is useful in deriving the bank's optimal choice of R:

LEMMA 6.1. *Recall that R^{**} is the argmax of $\mu_R - (c/K)F(R)$. On any interval of loan rates below R^{**} on which $r_b^m(R)$ is constant, the bank's constrained objective function* (6.3) *is strictly increasing in the loan rate R.*

The bank's "loan revenue"—expected payments from firms less the bank's cost of monitoring firms—is an increasing function of R when R is less than R^{**}. Increasing R causes a first-order stochastic dominance shift in bank portfolio returns, so when $r_b^m(R)$ is constant in R, D^m can be decreased. Thus, a higher loan rate increases loan revenue and decreases deposit payments, increasing bank profits.

Recall now that R^o is the lowest direct loan rate that gives investors a return of β. For R less than $R^{f(-1)}(R^o)$, $r_b^m(R)$ is constant at β, so Lemma 6.1 implies that the bank's expected profits increase in R for R less than $R^{f(-1)}(R^o)$. As a result, a monopoly bank will never choose R less than $R^{f(-1)}(R^o)$. This leads to the following result on a monopoly bank's ability to overcome autarky.

PROPOSITION 6.2. *Suppose the economy's direct-lending equilibrium is autarkic, so that either $\pi_f(R^o) < \beta$ or $r(R^*) < \beta$. A monopoly bank cannot overcome autarky if either* (i) $\pi_f(R^o) < \beta$, *or* (ii) $r(R^*) > \beta$ *and* $\pi_f(R^{**}) < \beta$.

Autarky occurs with direct lending either because investors will only lend on terms that are unattractive to firms or because no loan is attractive to investors. In the first case, since the bank never offers better lending terms than investors, firms earn less than autarkic returns, and so no one will become a firm in the first place. In the second case, a monopoly bank can always attract depositors by offering them their autarkic return β. Lemma 6.1 implies that the bank sets its loan rate at or above R^{**}; the second condition in (ii) then implies that this leaves firms with less than autarkic returns.

Comparing Corollary 5.3 and Proposition 6.2, the circumstances under which a monopoly bank overcomes autarky are more limited than those under which free-entry banking can do so. A key point here is that, by the time lending occurs, firms have incurred a sunk cost, so that if the bank exploits firms by raising its loan rate, it may push returns below the ex ante autarkic return β.[21]

As already noted, the main reason for allowing monopoly banking is to assure coordination when the benefits of diversification outweigh the benefits of capitalization, which suggests the economy is reasonably large. The next two propositions examine the behavior of a monopoly bank in such an economy. Since Proposition 6.2 dealt with the case where direct lending leads to autarky, these propositions assume direct lending is viable. Let the direct-lending equilibrium rate and number of firms be R^{d} and N^{d}, respectively, so that $n^{\mathrm{d}} = (M - N^{\mathrm{d}})/K$ is the number of funded firms under direct lending.

PROPOSITION 6.3. *Suppose the direct-lending equilibrium is nonrationed, so that $n^{\mathrm{d}} = N^{\mathrm{d}}$ and $\pi_{\mathrm{f}}(R^{\mathrm{d}}) = r(R^{\mathrm{d}})$. Then if n^{d} is sufficiently large, a monopoly bank will sustain less than n^{d} productive firms in equilibrium.*

The intuition is that, as long as there is no rationing, a sufficiently large bank will choose a rate greater than R^*.[22] Since this exceeds the direct-lending equilibrium rate R^{d}, firms are worse off than depositors. In equilibrium, there must be a smaller number of funded firms.[23]

If the direct-lending equilibrium involves rationing, the situation is more complex. Because the monopoly bank can profitably set its loan rate R above R^*, it can reduce (funded) firm profits relative to depositor returns; this makes being a firm less attractive and may reduce rationing. However, a large monopoly bank tends to set its loan rate above $R^{**} > R^*$, which may reduce funded firm profits so much that too few agents choose to become firms. Proposition 6.4 provides the conditions under which these outcomes hold.

PROPOSITION 6.4. *Suppose the direct-lending equilibrium is rationed, so that the number of funded firms n^{d} is less than N^{d}. Then* (i) *if $\pi_{\mathrm{f}}(R^{**}) \geq r(R^*)$, a monopoly bank in a sufficiently large economy can reduce rationing and increase the number of funded firms relative to direct lending.*

[21] Even if firms only had to expend effort in order to acquire the production technology, similar results would hold. Now the bank would exploit firms until they had not recouped any of their sunk cost.

[22] As shown in the Proof, if there is no rationing, the bank's profits are equal to the difference between total monitoring costs for direct lending at R^{f} and total monitoring costs under intermediation. If the bank is large, the depositors' costs of monitoring the bank are relatively small, and the effect of R on this component of costs is negligible. The main effect of increasing R is to increase the "wedge" between the direct cost $KcF(R)$ of monitoring each firm and the bank's cost $cF(R)$; the bank maximizes this by increasing R.

[23] While it is tempting to assume that there would be more depositors than there are firms for them to fund, this need not be the only equilibrium. However, having excess numbers of firms relative to depositors also shrinks the size of the bank, since the total number of agents in the economy is fixed.

However, (ii) *if* $\pi_f(R^{**}) < r(R^*)$, *the monopoly bank will reduce the number of funded firms relative to direct lending.*

I now consider the bank's incentive to freely limit its size. This depends in part on what happens when both depositors *and* firms are rationed: if rationed agents revert to autarky, the bank has less incentive to reduce its size than it does if rationed agents can recontract among themselves.[24] In either case, if increased capitalization outweighs reduced diversification, a reduction in bank size reduces a nonrationed depositor's cost of monitoring the bank. Although this would suggest that the bank can lower its deposit rate, depositors and firms will demand some compensation for their increased chance of being rationed. To the extent a lower deposit rate is possible, the bank might prefer a small reduction in size in exchange for an increased profit margin. Based on earlier discussion, this incentive should increase if there is unobservable systematic risk in the economy.

7. CONCLUSION

This paper has addressed bank behavior and industry structure in a finite delegated monitoring economy where the numbers of banks, firms, and depositors are endogenous. I have focused on three issues: the nature of a bank in this setting, equilibrium when free entry into banking is allowed, and equilibrium under monopoly banking.

In contrast to earlier research, I find that the case for a natural banking monopoly may be overstated. Since bank capital facilitates intermediation, increased bank size is not necessarily Pareto-improving, and a system with many banks may dominate one with a few large banks. Furthermore, multiple equilibria and coordination problems may make it difficult for an economy to achieve an optimal banking structure.

Nevertheless, government intervention to try to overcome coordination problems may be counterproductive. In particular, a monopoly bank causes distortions due to its desire to exploit firms, restricting output and in some cases failing to overcome autarky even when banking is ex ante feasible. In fact, the case where monopoly banking is least likely to distort output is when direct lending is feasible but would result in credit rationing. This in turn requires that monitoring costs be high enough so that credit rationing becomes an issue, but not so high that direct lending is infeasible. To the extent that monitoring costs decrease with an economy's stage of development, this suggests that the case for government

[24] Allowing agents to recontract makes it less costly to be rationed, which makes agents more willing to use the bank in the first place.

restrictions on entry into banking may be limited to economies of intermediate development, and even then is uncertain.

Because multiple equilibria and adoption externalities are common in theories of financial intermediation, further research on beliefs and their effect on strategic competition among banks would be useful. A dynamic setting would be useful, since banks have some durability and individuals' beliefs are tempered by experience. Winton (1994) addresses some of these issues using a more stylized model of competition among banks for investors.

A number of related areas remain to be studied. Further work on the tradeoffs between capital and diversification in the presence of nonobservable systematic and sectoral risk would have implications both for the historical evolution of banking systems and for public policy. In addition, most of the delegated monitoring literature assumes that an individual can monitor additional loans at constant marginal cost. Introducing increasing marginal costs of monitoring and allowing for teams and hierarchies of monitors could produce useful insights regarding the optimal size and structure of the banking firm. Finally, since interbank borrowing and lending are in fact used to resolve bank imbalances, the effect of these activities on equilibrium structure should be examined.

APPENDIX

Proof of Proposition 3.1. From the text, once agents have chosen whether or not to acquire the technology, loan rates are R^* if there are more firms than investors can finance ($N^d > (M - N^d)/K$), R^o if there are more investable funds than firms can use ($N^d < (M - N^d)/K$), and any rate in the interval $[R^o, R^*]$ otherwise. The Proposition follows from examining which of these outcomes are consistent with agents' initial decisions to become firms or investors.

(i) $r(R^*) < \beta$ or $\pi_f(R^o) < \beta$. If $r(R^*) < \beta$, consuming one's own endowment is always better than investing in a firm, so no investor is willing to lend; thus firms earn zero profits. If $\pi_f(R^o) < \beta$, then any rate capable of attracting investors leaves firms with a lower return than they could have received by consuming their endowment. In both cases, no one is willing to become a firm in the first place.

(ii) $\beta \leq r(R^*) < \pi_f(R^*)$. Suppose that there are more investable funds than firms can use ($N^d < (M - N^d)/K$). Since the loan rate can't exceed R^*, and all firms receive funding for sure, investors' expected returns are strictly lower than those of firms, and at least one investor has incentive to deviate to becoming a firm. If $N^d = (M - N^d)/K = n^d$, the same logic holds, so long as the number of agents in the economy is large enough that

$(n^d - 1)/(N^d + 1)$ times $\pi_f(R^*)$ isn't below $r(R^*)$. Thus, subject to the "integer effects" discussed in Footnote 11, the only equilibrium involves more firms than can be funded ($N^d > (M - N^d)/K = n^d$), such that $r(R^*) \approx (n^d/N^d)\pi_f(R^*)$.

(iii) $\beta \leq \pi_f(R^o) \leq \pi_f(R^*) \leq r(R^*)$. Suppose that too many agents become firms, so that $N^d > (M - N^d)/K$ and the loan rate is R^*. Then $r(R^*) > (n^d/N^d)\pi_f(R^*)$. Investors are strictly better off than firms, and at least one firm would prefer to deviate and remain an investor in the first stage.

Suppose instead that there are more investable funds than firms can use, so $N^d < (M - N^d)/K$, and the loan rate is R^o. Unless $\pi_f(R^o) = \beta$ (which by definition equals $r(R^o)$), investors are strictly worse off than firms, and at least one investor prefers to deviate and become a firm at the first stage.

Thus, unless $\pi_f(R^o) = \beta$, the only equilibrium outcome is for there to be just enough firms to absorb all investable funds ($N^d = (M - N^d)/K$). To be consistent with equilibrium, R must satisfy $r(R) \approx \pi_f(R)$; otherwise, investors or firms would have incentive to switch their initial choice of role. Since $\pi_f(R^*) > r(R^*)$, there is a rate between R^o and R^* for which this holds.

Q.E.D.

Proof of Proposition 4.2. Suppose the bank lends to one firm at rate R. The bank's portfolio return is $\tilde{W} = K \cdot \min\{\tilde{X}, R\}$, which has distribution function $F(w/K)$ for $w < KR$, and 1 otherwise. If depositors receive a debt contract with stated rate D, they receive $\min\{\tilde{W}/(K - 1), D\}$ and monitor whenever $\tilde{W} < (K - 1)D$. Equations (3.2) and (3.3) imply that each depositor's expected return $r_b(R, D; 1)$ is

$$r_b(R, D; 1) = D - \int_0^D F[y(K - 1)/K]\,dy - cF[D(K - 1)/K]$$

$$= D - \frac{K}{K - 1} \cdot \int_0^{D(K-1)/K} F(y)\,dy - cF[D(K - 1)/K], \quad \text{(A.1)}$$

so long as $D < KR/(K - 1)$. The bank receives expected payment $K\mu_R$ from its loan, incurs monitoring cost $cF(R)$, and pays each of $K - 1$ depositors the first two terms on the RHS of (A.1), so that bank's expected profits $\pi_b(R, D; 1)$ are

$$\pi_b(R, D; 1) = K\mu_R - (K - 1)D + K \cdot \int_0^{D(K-1)/K} F(y)\,dy - cF(R). \quad \text{(A.2)}$$

If D is chosen to equate $r_b(\cdot)$ and $\pi_b(\cdot)$, both bank and depositors prefer intermediation to direct lending at R. To see this, note that $r_b = \pi_b$ implies

$$\mu_R - D + \frac{K}{K-1} \cdot \int_0^{D(K-1)/K} F(y)\, dy = (c/K) \cdot [F(R) - F[D(K-1)/K]]. \tag{A.3}$$

Recall that the direct lending return $r(R) = \mu_R - cF(R)$. Using this and (A.1), it follows that $r_b(\cdot) > r(R)$ iff [LHS of (A.3)] $< cF(R) - cF[D(K - 1)/K]$. Since the latter expression is positive and equals K times the RHS of (A.3), it follows that $r_b(\cdot)$ and $\pi_b(\cdot)$ exceed $r(R)$ when $r_b(\cdot) = \pi_b(\cdot)$.

For $D < KR/(K-1)$, both $\pi_b(\cdot)$ and $r_b(\cdot)$ are continuous in D and $\pi_b(\cdot)$ is decreasing in D. At $D = 0$, $\pi_b(\cdot) > 0$ and $r_b(\cdot) = 0$; as $D \to KR/(K-1)$ from below, $\pi_b(\cdot) < 0$ and $r_b(\cdot) > 0$. Thus equality holds for some intermediate D.

Q.E.D.

Lemma A.1. (i) *$H_n(w; R)$ has probability mass $(1 - F(R))^n$ concentrated at nKR and is continuous elsewhere.* (ii) *$h_n(w; R)$ has discontinuities at all integral multiples of KR; it is well defined and continuous at all other points in* $[0, nKR]$. (iii) *If $X_m \geq R > R' \geq 0$, then for any number of bank loans n, $H_n(w; R)$ has first-order stochastic dominance over $H_n(w; R')$.* (iv) *The same statements apply to G_n and g_n, with K replaced by $K/(nK-1)$.*

Proof of Lemma A.1. (i) Because the variables $K \cdot \min(\tilde{X}_i, R)$ are mutually stochastically independent, H_n is the n-fold convolution of $K \cdot \min(\tilde{X}_i, R)$'s distribution. The latter is given by

$$F_R(Kx) = \begin{cases} 0, & \text{if } x \leq 0; \\ F(x), & \text{if } x \in (0, R); \\ 1, & \text{if } x \geq R. \end{cases} \tag{A.4}$$

F_R is continuous everywhere except at KR, where it has an atom with mass $1 - F(R)$. The result follows immediately: see Feller (1971, p. 147). (ii) Piecewise continuity of h_n follows because, within each interval of the form $(iK, (i+1)K)$, at most i firms can be paying their debts in full, while at $(i+1)K$ another such finite probability event becomes possible. (iii) Whenever $R' < R$, $F_{R'}(Kx) \geq F_R(Kx)$ for all x, with strict inequality for x in $[R, R')$. Thus, F_R has first-order stochastic dominance over $F_{R'}$. This carries over to any number of convolutions of each distribution: see Winton (1990, Chapter 3, Lemma B.2). (iv) This follows because $G_n(y; R) = H_n[(nK-1)y; R]$.

Q.E.D.

Proof of Lemma 4.3. Part (i) is obvious. Part (ii) follows because an increase in R causes a first-order stochastic dominance shift in bank portfolio returns, improving expected depositor cash flows and reducing the probability that depositors must monitor. Part (iii) follows from similar considerations.

Q.E.D.

Proof of Lemma 4.4. (i) and (ii) follow directly from Lemma 4.3 and expressions (4.1) and (4.2). For (iii), note that $\partial r_b(R, D; n)/\partial D = 1 - G_n(R, D; n) - cg_n(R, D; n)$. When $n = 1$, this equals $1 - F[D(K - 1)/K] - cf[D(K - 1)/K] \cdot (K - 1)/K$, which exceeds $1 - F(R^*) - cf(R^*) = 0$ when the conditions of the Lemma hold. When n is large and D is less than μ_R, the last term in the expression for the derivative is negligible compared with $1 - G_n(\cdot)$, which is positive and converges to 1.

Q.E.D.

Proof of Lemma 4.5. After some manipulation, $\pi_b(R, D; n') - \pi_b(R, D; n)$ becomes

$$\frac{n' - n}{n} \cdot \left[\pi_b(R, D; n) - D + \int_0^D G_n(y; R)\, dy\right] + (n'K - 1) \cdot \int_0^D [G_{n'}(y; R) - G_n(y; R)]\, dy. \tag{A.5}$$

The first bracketed term is the difference between bank profits and expected payments to one depositor. When $\pi_b = r_b$, this equals $-cG_n(D; R)$. The second integral is the negative of the change in expected payments to a depositor caused by the increase in bank size. The rest of the Lemma follows easily.

Q.E.D.

Proof of Corollary 5.3. (i) Under direct lending, autarky occurs when $\beta > \pi_f(R^o)$ or $\beta > r(R^*)$. Since an FE with only one loan per bank is always possible and strictly Pareto-dominates direct lending, continuity arguments imply that there are economies for which direct lending is infeasible but an FE exists.

(ii) This follows from Propositions 4.1 and 5.2. Suppose the economy is large enough so that an FE with large banks exists. Since $r_b(\cdot) = \pi_b(\cdot)$ in such an equilibrium, (4.1) and (4.2) imply that

$$\mu_R - D + \int_0^D G_n(y; R)\, dy = (c/K) \cdot [F(R) - G_n(D; R)/n], \tag{A.6}$$

so that

$$r_b(R, D; n) = \pi_b(R, D; n) = \mu_R - (c/K) \cdot F(R) - \frac{nK - 1}{nK} \cdot cG_n(D; R). \tag{A.7}$$

For n sufficiently large and D less than μ_R, $G_n(D; R) \to 0$. Now choose $R = R'$, and choose D so that (A.6) and (A.7) hold. Then (A.7) is less than β by the last term, which is arbitrarily small. Since $R' < R^{**}$, increasing R by $\varepsilon > 0$ increases the first two terms of (A.7) by a finite amount, while the change in $G_n(\cdot)$ is arbitrarily small. Thus there is a small $\varepsilon > 0$, such that, for $R = R' + \varepsilon$ and appropriate choice of D, (A.6) holds and $r_b(\cdot) = \pi_b(\cdot) = \beta$. Also, ε small and $\pi_f(R') > \beta$ imply $\pi_f(R' + \varepsilon) > \beta$. Now increase ε; either there is an R (and D) for which bank, depositor, and firm returns are equated above β, or else firm profits are greater than bank and depositor returns. The first case leads a nonrationed FE, while the second case leads to a rationed FE.

Q.E.D.

Proof of Proposition 5.4. (i) When $R^e > R^d$, $\pi_f(R^e) < \pi_f(R^d)$. If $B^e \cdot n^e < n^d$, then $N^e > N^d$, so $(B^e \cdot n^e/N^e) \cdot \pi_f(R^e) < (n^d/N^d) \cdot \pi_f(R^d)$. But since the FE cannot be dominated by a direct loan, $r_b(R^e, D^e; n^e) \geq r(R^d)$. But $(B^e \cdot n^e/N^e) \cdot \pi_f(R^e) = r_b(R^e, D^e; n^e)$, and $(n^d/N^d) \cdot \pi_f(R^d) = r(R^d)$, which is a contradiction. (ii) The same logic does not necessarily lead to a contradiction in the case when $R^e \leq R^d$, so it is possible that rationing increases in this case.

Q.E.D.

Proof of Proposition 5.5. (i) Since credit rationing is increased, Proposition 5.4 implies $R^e \leq R^d$. From (4.1) and (4.2), the total welfare of a bank and its depositors equals $nK\mu_R - nc \cdot F(R) - (nK - 1)c \cdot G_n(D^e; R^e)$, which increases with R for $R \leq R^{**}$. Since $R^{**} > R^d$, it is feasible for a bank to increase its loan rate R slightly from R^e, improving profits of deviating firms from $(B^e \cdot n^e/N^e) \cdot \pi_f(R^e)$ to $\pi_f(R)$ and improving depositor returns by Lemma 4.3. If the bank's expected return increases, the Proof is done. If not, decrease D until $\pi_b(R, D; n) = \pi_b(R^e, D^e; n) = r_b(R^e, D^e; n)$. This increases the total welfare of the bank and its depositors; since the bank's expected profits are unchanged from the initial equilibrium, depositor's expected returns must be strictly higher than $r_b(R^e, D^e; n)$.

(ii) Suppose the bank sets its loan rate $R = R^e + \varepsilon$; as shown in (i), for small ε, this makes deviating firms better off. Also, $\Delta\pi_b \equiv \pi_b(R, D^e; n^e) - \pi_b(R^e, D^e; n^e) \approx \varepsilon \cdot [\partial\pi_b/\partial R]$, where

$$\frac{\partial \pi_b}{\partial R} = nK \cdot [1 - F(R^e) - (c/K) \cdot f(R^e)] + (nK - 1) \cdot \int_0^{D^e} \frac{\partial G_n(y; R^e)}{\partial R} \cdot dy. \tag{A.8}$$

From (i), $R^e < R^{**}$, so the bracketed term in (A.8) is always positive. (a) For $n = 1$, $G_n(D^e; R^e) = F[D^e(K - 1)/K]$. So long as $D^e < KR^e/(K - 1)$ (which must hold in an FE), $\partial G_n(\cdot)/\partial R = 0$. Thus $\partial \pi_b/\partial R > 0$, so $\Delta \pi_b > 0$ and the bank is better off. (b) For large n and $D^e < \mu_R$, G_n is small for all $y < D^e$, and the integral term in (A.8) is arbitrarily small. Thus $\partial \pi_b/\partial R > 0$, so $\Delta \pi_b > 0$.

Q.E.D.

Proof of Lemma 6.1. Using (6.3) and noting that $r_b^m(R)$ is constant, one has

$$\frac{d\pi_b}{dR} = nK \cdot [1 - F(R) - (c/K) \cdot f(R)] - (nK - 1)c \cdot \left[g_n(D^m; R) \cdot \frac{dD^m}{dR} + \frac{\partial G_n(D^m; \mathrm{R})}{\partial R}\right]. \tag{A.9}$$

Since $\mu_R - cF(R)$ is quasiconcave, the first bracketed term is positive for any $R < R^{**}$. $\partial G_n(\cdot)/\partial R$ is negative, so if $dD^m/dR < 0$, (A.9) is positive as claimed. Applying the Implicit Function Theorem to (6.2) yields

$$\frac{dD^m}{dR} = \frac{\partial G_n(D^m; \mathrm{R})/\partial R}{1 - G_n(D^m; \mathrm{R}) - cg_n(D^m; R)}. \tag{A.10}$$

The numerator of (A.10) is negative, while the denominator is positive (if not, both the bank and its depositors would benefit from a decrease in D, violating the definition of D^m as the *smallest* rate satisfying (6.2)). Thus, $dD^m/dR < 0$.

Q.E.D.

Proof of Proposition 6.2. (i) If $\pi_f(R^o) < \beta$, any rate that attracts direct lending leaves firms with less than autarkic returns. As noted in the text, the bank never lends at less than $R^{f(-1)}(R^o)$; by the definition of R^f, firms' expected profits are less than $\pi^f(R^o)$. Thus no one will choose to become a firm in the first place. (ii) Now direct lending is never attractive, so the monopoly bank sets $r_b^m(R) = \beta$. Since this is constant, Lemma 6.1 implies that the bank never chooses R less than R^{**}. When $\pi_f(R^{**}) < \beta$, firms are left with worse than autarkic returns, so no one will choose to become a firm.

Q.E.D.

Proof of Proposition 6.3. Suppose that the number of firms in the monopoly economy is n^{d}. Since the direct-lending equilibrium is nonrationed, there are exactly enough depositors to fund all firms. I show below that, if n^{d} is large, the bank chooses a loan rate $R > R^{\mathrm{d}}$, which means firm profits are less than depositor returns: $\pi_{\mathrm{f}}(R) < \pi_{\mathrm{f}}(R^{\mathrm{d}})$, while $r_{\mathrm{b}}^{\mathrm{m}}(R) \geq r(\min\{R^{\mathrm{f}}(R), R^*\}) = r(\min\{R, R^*\}) \geq r(R^{\mathrm{d}}) = \pi_{\mathrm{f}}(R^{\mathrm{d}})$. Thus n^{d} cannot be an equilibrium number of firms. If more or fewer agents choose to be firms, fewer than n^{d} firms are funded in equilibrium.

Now I show $R > R^{\mathrm{d}}$ when there are n^{d} firms. From (6.3), bank profits are

$$nK \cdot [\mu_R - (c/K)F(R)] - (nK - 1) \cdot [r(\min\{R, R^*\}) + cG_n(D^{\mathrm{m}}; R), \tag{A.11}$$

where the superscript on n^{d} is suppressed. If $R \leq R^*$, (A.11) becomes

$$\mu_R + (nK - n - 1) \cdot cF(R) - (nK - 1) \cdot cG_n(D^{\mathrm{m}}; R). \tag{A.12}$$

If R increases, μ_R and $F(R)$ increase. For n large, $G_n(D_{\mathrm{m}}; R)$ is arbitrarily small for all R, so the change in G_n from the increase in R is small compared to the change in $F(R)$, and so bank profits increase. Thus, the bank chooses R arbitrarily close to or greater than R^*. In fact, by Lemma 6.1, since $r_{\mathrm{b}}^{\mathrm{m}}(R) = r(R^*)$ for all $R > R^*$, bank profits increase in R on $[R^*, R^{**}]$, and the bank chooses $R \geq R^{**}$. If $R^{\mathrm{d}} < R^*$, this shows that the bank chooses R greater than R^{d}. If $R^{\mathrm{d}} = R^*$, the fact that firm profits are positive in the direct-lending equilibrium implies that R^* is an interior point. Thus $R^{**} > R^*$, and the bank chooses $R > R^* = R^{\mathrm{d}}$.

Q.E.D.

Proof of Proposition 6.4. (i) If the number of firms in the monopoly economy is N and $n \leq N$ are funded, where n is large and $n = (M - N)/K$ (no depositor rationing), bank profits are

$$nK \cdot [\mu_R - (c/K)F(R)] - (nK - 1) \cdot [r(\min\{R^{\mathrm{f}}(R), R^*\}) + cG_n(D^{\mathrm{m}}; R)]. \tag{A.13}$$

If $N = n$, $R^{\mathrm{f}}(R) = R$, and the proof of Proposition 6.3 shows that the bank optimally sets $R \geq R^{**}$. If $N > n$, $R^{\mathrm{f}}(R) > R$. If the bank chooses R such that $R^{\mathrm{f}}(R) \geq R^*$, (A.13) is increasing in R up to R^{**}, and the bank should choose $R \geq R^{**}$. If the bank chooses R such that $R^{\mathrm{f}}(R) < R^*$, its profits are less than they would be if R^{f} in (A.13) were replaced by R. (A.13) with R^{f} replaced by R is the same as (A.11), and so the bank is better off choosing $R \geq R^{**}$.

Both bracketed terms in (A.13) decrease in R for $R > R^{**}$. However, for large n the change in the second term is negligible relative to the change in the first term; thus the bank will not choose R much greater than R^{**}. If $\pi_f(R^{**}) = r(R^*)$, then since $r_b^m(R^{**}) = r(R^*)$, depositors and firms earn the same return and the bank's choice of R is consistent with a nonrationed equilibrium. If $\pi_f(R^{**}) > r(R^*)$, then the bank's choice of R is consistent with a rationed equilibrium. Since direct-lending equilibrium is rationed, $(n^d/N^d) \cdot \pi_f(R^*) = r(R^*)$. $R^{**} \geq R^*$ implies $\pi_f(R^{**}) \leq \pi_f(R^*)$; since monopoly equilibrium requires $(n/N) \cdot \pi_f(R^*) = r_b^m(R^{**}) = r(R^*)$, it follows that n/N is greater than n^d/N^d. Thus, the monopoly bank can reduce rationing when $\pi_f(R^{**}) \geq r(R^*)$.

(ii) The same steps used in (i) show that if the number of firms in the monopoly economy is N and $n \leq N$ are funded, where n is large and $n = (M - N)/K$, the bank would choose $R \approx R^{**}$. Since $\pi_f(R^{**}) \leq r(R^*)$, and firms may be rationed, firms would prefer to be depositors. This cannot be an equilibrium, so either n is not large or $n = N < (M - N)/K$ (depositors are rationed).

If n is large and depositors are rationed, the bank can either set $R < R^*$ or $R \geq R^*$. If $R < R^*$, $r_b^m(R) < r^*$, and firm profits $\pi_f(R) > \pi_f(R^*)$. Since $\pi_f(R^*) > r(R^*)$, firms are better off than depositors, and depositor rationing is inconsistent with equilibrium. If $R \geq R^*$, $r_b^m(R)$ is constant, so by Lemma 6.1 the bank chooses $R \geq R^{**}$. This leaves depositors with expected return $r_b^m(R) = r(R^*)$, which exceeds firm profits $\pi_f(R^{**})$. Thus, although depositors are rationed, depositors are still better off than firms, and this cannot be an equilibrium. Equilibrium must involve a relatively small number of funded firms.

Q.E.D.

REFERENCES

Bhattacharya, S., and Thakor, A. (1993). Contemporary banking theory, *J. Finan. Intermediation* **3**, 2–50.

Barlow, R., and Proschan, F. (1975). "Statistical Theory of Reliability and Life Testing," Holt, Rinehart & Winston, New York.

Baumol, W., Panzar, J., and Willig, R. (1985). On the theory of perfectly contestable markets, *in* "New Developments in the Analysis of Market Structure" (J. Stiglitz and F. Mathewson, Eds.), pp. 339–365. MIT Press, Cambridge, MA.

Boyd, J., and Prescott, E. (1986). Financial intermediary coalitions, *J. Econ. Theory* **38**, 211–232.

Daltung, S. (1991). "Optimal Structure of the Financial Intermediation Industry," Seminar Paper No. 489, Institute for International Economic Studies, Stockholm University.

Diamond, D. (1984). Financial intermediation and delegated monitoring, *Rev. Econ. Stud.* **51**, 393–414.

FELLER, W. (1971). "An Introduction To Probability Theory And Its Applications," Vol. II, 2nd ed., Wiley, New York.

GALE, D., AND HELLWIG, M. (1985). Incentive compatible debt contracts: The one period problem, *Rev. Econ. Stud.* **52,** 647–663.

KRASA, S., AND KUBITSCHEK, J. (1988). Direct Financing, Intermediation and Credit Rationing, working paper, Stanford University.

KRASA, S., AND VILLAMIL, A. (1992a). Monitoring the monitor: An incentive structure for a financial intermediary, *J. Econ. Theory* **57,** 197–221.

KRASA, S., AND VILLAMIL, A. (1992b). A theory of optimal bank size, *Oxford Econ. Pap.* **44,** 725–749.

MOOKHERJEE, D., AND PNG, I. (1989). Optimal auditing, insurance, and redistribution, *Quart. J. Econ.* **104,** 399–415.

MYERS, S. (1977). Determinants of corporate borrowing, *J. Finan. Econ.* **5,** 147–175.

ROTHSCHILD, M., AND STIGLITZ, J. (1970). Increasing risk I: A definition, *J. Econ. Theory* **2,** 225–243.

SCHULER, K. (1992). The world history of free banking: An overview, *in* "The Experience of Free Banking" (K. Dowd, Ed.), pp. 7–47. Routledge, London.

SMITH, C. (1986). Raising capital: Theory and evidence, *J. Appl. Corp. Finance* **4,** 6–21.

TOWNSEND, R. (1979). Optimal contracts and competitive markets with costly state verification, *J. Econ. Theory* **21,** 265–293.

WILLIAMSON, S. (1986). Costly monitoring, financial intermediation, and equilibrium credit rationing, *J. Monet. Econ.* **18,** 159–179.

WINTON, A. (1990). "Three Essays on Information, Contracting, and Financial Intermediation." Ph.D. Dissertation, University of Pennsylvania.

WINTON, A. (1994). "Competition Among Financial Intermediaries When Diversification Matters," Banking Research Center working paper 185, Northwestern University.

WINTON, A. (1995). Costly state verification and multiple investors: The role of seniority, *Rev. Finan. Stud.* **8,** 91–123.

YANELLE, M. (1988). "On the Theory of Intermediation," Unpublished Ph.D. Dissertation, Rheinischen Friedrich-Wilhelms Universität zu Bonn.

YANELLE, M. (1989). The strategic analysis of intermediation, *Europ. Econ. Rev.* **33,** 294–301.