

Competition among Financial Intermediaries When Diversification Matters*

Andrew Winton

Finance Department, Northwestern University, Evanston, IL 60208–2001

Received December 2, 1996

Many models predict that the diversification and efficiency of financial intermediaries (“banks”) increases with their size, so that a relatively unrestricted banking sector will settle into an equilibrium with several large, well-diversified, and competitive banks. However, this prediction is at odds with the actual pattern of unrestricted banking sector evolution in many countries. I develop a model that motivates this actual pattern and examine the model’s implications for regulatory policy. I show that an investor’s return from a bank depends on the number of investors using that bank; this adoption externality makes investor beliefs about other investors’ actions critical for bank competition. In a young banking system with free entry, coordination problems lead to excessive fragmentation, and debt overhang makes it difficult for small banks to capture additional market share. As the system matures, many banks fail, and the survivors become the focus of investor beliefs; these incumbents gain a strong advantage over entrants, facilitating collusion. Entry restrictions reduce fragmentation but aid collusion, while government insurance for investors reduces incumbency advantage and collusion but may cause excessive fragmentation. Thus, regulators may wish to impose temporary entry restrictions, along with partial insurance. These results are consistent with historical evidence from several countries. *Journal of Economic Literature* Classification Numbers: G21, G22, L13. © 1997 Academic Press

1. INTRODUCTION

Many models of financial intermediaries (e.g., Diamond, 1984, Ramakrishnan and Thakor, 1984) imply that these institutions have substantial

* This paper originally formed Chapter Four of my Ph.D. dissertation, which was supported by a Unisys Fellowship. Since then, I have benefited from a research grant from the Banking Research Center at Northwestern University. I thank David Besanko, Susan Chaplinsky, Mike Fishman, Thomas Gehrig, Stuart Greenbaum, Laurie Hodrick, Rich Kihlstrom, Raghu Rajan, Anjan Thakor (the Editor), Anne Villamil, Asher Wolinsky, and two anonymous referees for their comments, along with the members of the Penn Shadow Theory Workshop and participants at the *Journal of Financial Intermediation* Symposium on the Reform of Financial Institutions and Markets. All mistakes remain my responsibility.

scale economies linked to improved diversification. History provides some support for this view. For example, in the period leading up to the Great Depression, the U.S., which restricted most banks to a single state, had higher bank failure rates than Canada, which had large, nationwide banks; within the U.S., larger banks were less likely to fail than smaller banks (Calomiris, 1993). Some have used this to argue that, if regulators would only leave intermediaries alone, free entry and competition would quickly lead to the dominance of several large, diversified, competitive intermediaries, laying to rest concerns about safety and collusion alike.

A closer look at the historical record suggests that this conclusion may be too sanguine. Consider the experience of Australia, Canada, and Scotland, where the formation of intermediaries was relatively unrestricted. In each case, one sees high initial entry, slow consolidation through failure and merger, and a strong tendency to collusion among the survivors, even when these numbered a dozen or more. Once consolidation occurred, these systems seem to have been almost immune to entry by new institutions. By contrast, the more fragmented banking system of the late 19th and early 20th century U.S. did allow significant entry, and entry was greatest in those states that had formal deposit insurance. What forces account for these patterns, and what are the implications of these forces for bank regulation?

In this paper, I suggest that these patterns can be explained by looking more closely at how diversification affects competition among financial intermediaries. Suppose that, because of fixed costs in providing financial services to a given sector or group, larger intermediaries (“banks”) are potentially better diversified than smaller ones, and that diversification improves the quality of the service the bank provides to investors. Furthermore, suppose that the bank funds itself with debt-like contracts, as do many intermediaries, such as banks, finance companies, and insurers. These features—common to many models of financial intermediation, such as delegated monitoring or insurance—have several implications for bank competition.

The link between size and diversification implies that the more investors use a given bank, the better off the investors are. This adoption externality makes investor beliefs self-fulfilling: if all believe a bank is small, hence risky and not worth investing in, few investors will go to that bank, and it will in fact be small, hence risky and not worth investing in. If a bank wants to establish a reputation for being large, diversified, and thus well worth investing in, it must first get large: i.e., it must attract more investors. The only short-run tool at the bank’s disposal is the “stated rate” (principal plus interest) that it offers investors. By raising this rate, it can offer a higher expected return to any investors already planning to use the bank; seeing this, marginal investors at other banks will now choose this bank,

making the bank larger and better diversified, and further improving investors' expected return.

However, raising its stated rate is costly to the bank for two reasons. First, even without a change in size, the higher stated rate clearly decreases bank profits. Second, if the bank's size does increase, then, even for a fixed stated rate, improved diversification increases the bank's expected payments to each investor: debtholders benefit from reduced risk. As a result, if the bank's initial profit margins are not too high, it may prefer not to try to increase its market share. This is a form of the Myers (1977) underinvestment or "debt overhang" problem, where new business that reduces overall risk may not be chosen because it benefits debtholders at the expense of shareholders. This debt overhang effect weakens rate competition among banks, promoting collusion.

This apparent contradiction—banks may not wish to increase size despite economies of scale from diversification—is rooted in the adoption externality already mentioned. Although a larger, better diversified bank can give investors a better expected return while offering a lower stated rate than smaller, riskier rivals, getting investors to accept the lower stated rate relies on the bank having a reputation for being larger and better-diversified. To get this reputation, the bank must attract and maintain a larger market share for some period of time, which is costly. The less transparent the bank's situation and the less stable the banking system, the longer the bank will have to maintain a large market share in order to change its reputation, making the *eventual* improvement in profits less attractive relative to the immediate costs.

To see how this affects the evolution of intermediary structure, first consider a market where banks are not yet well established, either because intermediation is new or because market growth is high. In such a setting, it is unlikely that investor beliefs will have coordinated on any particular bank, so banks that enter split up the market. Here, the reverse of the debt overhang effect suggests that many banks will enter: for any given stated rate, banks have higher expected profit margins as they are smaller. Indeed, there is a tendency for the maximum number of (barely viable) banks to enter.

As time goes on, many of these small, risky banks will eventually fail. Investor beliefs are likely to focus on the survivors, giving them an incumbency advantage over new entrants: if, all else equal, investors expect incumbents to have a larger market share and thus greater diversification, any given stated rate from an incumbent offers investors a higher expected total return than the same rate from a new entrant. Thus, investor beliefs become a barrier to entry. The debt overhang cost of expanding through rate competition may make collusion among the incumbent banks feasible even when their numbers are somewhat large.

Now consider the problem of regulators maximizing social welfare. Encouraging free entry in the early stages of financial intermediation may lead to many inefficiently small intermediaries, with high rates of failure and ensuing costs. However, although restricting initial entry should help focus investors on a relatively small number of larger (hence safer) intermediaries, this makes collusion more likely. Furthermore, as time passes, investor beliefs are likely to focus more and more on the incumbents, making it hard for “de novo” entrants to mount a credible threat even if regulators later drop entry restrictions. If banks in other jurisdictions are already well diversified, these institutions might be able to enter—if investors in the “domestic” region are sufficiently aware of these “foreign” banks’ safety, and if the foreign banks can overcome the usual informational, cultural, and political difficulties of operating in a different region. In fact, as discussed in Section 7, it was the arrival of large foreign banks that finally broke the oligopoly of Britain’s large “clearing banks” in the 1960s.

As an alternative, regulators can insure investors; by reducing investors’ concern for the risk and thus the size of banks, such insurance weakens the advantage of incumbents, enhancing entry and competition. Indeed, the introduction of deposit insurance in Canada in 1967 seems to have had these effects (see Section 7). However, as U.S. experience shows, entry may be excessive, especially if guarantee fees are fixed regardless of risk. This suggests that regulators may wish to impose entry restrictions for some time, to prevent initial fragmentation; also, upon lifting such restrictions, they may not wish to fully insure investors, so that investors still give some preference to large incumbents.

These results rely critically on the assumptions that greater size is required for better diversification, that investors find it costly to coordinate their actions, and that investor beliefs are slow to change. These are most likely to apply in settings where information problems loom large, increasing the cost of risk-sharing agreements among institutions, coordination among investors, and credible disclosure of institutions’ risk levels to investors. Thus, both historically and currently, the model’s conclusions should apply most strongly to intermediation in developing economies, where such market imperfections are likely to play a greater role.

My analysis is related to several recent papers that explore competition among banks when diversification plays a role. Using the delegated monitoring paradigm of Diamond (1984), Yanelle (1989, 1997) explores the impact of different contracting structures, while Winton (1995) focuses on the role of bank capital and the choice between competitive and monopolistic regimes. Yosha (1997) examines efficient industry structure for risk-sharing intermediaries as a function of the size of the economy. Matutes and Vives (1996) examine spatial competition between two banks in the presence of

transportation costs and find that deposit insurance increases rate competition and bank failure risk. Gehrig (1995) shows that spatial rent-seeking causes excessive entry. In Besanko and Thakor (1993), insured banks that diversify forfeit gains from risk shifting but increase their odds of surviving to collect informational rents from lending relationships; free entry diminishes these rents and encourages risk shifting. My work differs from these papers by focusing on how investor beliefs interact with bank competition, entry, and the impact of regulation to influence intermediary industrial structure.

A second related line of research examines the influence of adoption externalities on firm competition. Katz and Shapiro (1985) show that, if customer utility increases with the supplier's market share, then supplier incentives to provide compatible products are less than social incentives, and this may result in inefficient fragmentation. Farrell and Saloner (1986) show that adoption externalities can prevent users from switching to a new product even if it is innately superior, because users are worse off if they are the only ones to use the new product. My paper resembles these papers in showing how adoption externalities can lead to inefficient fragmentation or excessive incumbency power. However, because the firms on which I focus are intermediaries whose "product" is a debt contract, increased market share has an immediate negative impact on firm profits; this strengthens the fragmentation result in new markets and undermines competition in mature markets.

The rest of the paper is organized as follows. Section 2 motivates and sets out the basic framework. Section 3 examines investor beliefs and equilibrium outcomes when multiple banks compete in a static setting. Sections 4 and 5 extend the analysis to a dynamic setting, where banks interact repeatedly and market shares in the present may influence investor beliefs and behavior in the future. Section 6 examines regulatory issues. Section 7 discusses historical evidence that is consistent with the paper's implications, and Section 8 concludes.

2. BASIC MODEL AND ASSUMPTIONS

A. *General Considerations*

Four key modeling features are required for my analysis: (i) financial intermediaries provide a service that improves with diversification, (ii) diversification tends to increase with intermediary size, (iii) intermediaries optimally issue debt or debt-like contracts to investors, and (iv) large numbers of investors cannot coordinate their actions. Before going through the model's specifics, I discuss the motivation for these more general features.

Financial Intermediation Improves with Diversification. In several models of financial intermediation there is a cost linked to intermediary risk, and a better diversified intermediary has less risk and thus lower costs. In models of insurance or liquidity provision, investors are risk averse and face some risk which the intermediary can pool and diversify on their behalf. In models of delegated investment monitoring or evaluation, the possibility of bad outcomes allows the intermediary to hide proceeds (if these are costly to verify) or to claim that bad luck rather than lack of effort led to the bad outcomes (if effort is unobservable); an intermediary with better diversified investments has less chance of very bad outcomes, reducing associated costs.¹ Real intermediaries may perform several of these functions: e.g., commercial banks provide liquidity and evaluate and monitor loans, and life insurers provide insurance and evaluate and monitor privately placed bonds and commercial mortgages.

Diversification Tends to Increase with Intermediary Size. If there are fixed costs to monitoring or evaluating an individual borrower, delegation to one or a few intermediaries avoids costly duplication of effort. By the same token, a larger intermediary can serve more borrowers, and so be better diversified, at lower cost. The same applies to diversifying across types of borrowers when there are fixed costs to developing skills to deal with each type of borrower, and to diversifying across customers in an insurance setting when there are fixed costs to providing insurance to a single customer or type of customer.

Of course, if intermediaries can credibly share information with ease, they can cheaply share risks among themselves without excessive agency problems, allowing diversification without an increase in scale. In reality, risk-sharing techniques such as loan sales and syndications among banks and reinsurance contracting among insurers do involve some costs and agency problems, but their existence suggests that my model's results apply most strongly to times and places where the technology for credibly sharing information among intermediaries is least developed, and vice versa.

A related issue is that, to the extent an intermediary has some choice over how and where it expands (in terms of asset risk in particular), the degree to which size improves diversification can be partly under the institution's control and is likely to be partly unobservable. As I argue in Section 5, this actually strengthens my model's results: a small and risky bank has the least incentive to expand in a more diversified manner, especially since

¹ For liquidity provision, see Diamond and Dybvig (1983), Chari and Jagannathan (1988), Jacklin and Bhattacharya (1988), and Gorton and Pennacchi (1990); for delegated investment evaluation, see Campbell and Kracaw (1980), Ramakrishnan and Thakor (1984), and Boyd and Prescott (1986); references for delegated investment monitoring are given below. For further references and discussion, see Bhattacharya and Thakor (1993).

it will take time to convince the investors that the bank is both expanding and choosing the less risky strategy.

Finally, there may be diseconomies of scale (e.g., from managing larger numbers of employees and managers), so that intermediary performance eventually begins to deteriorate with size. As I argue at the end of Section 3, this does not change my model's qualitative results so long as there is a significant range of bank sizes over which diversification is valuable. There is evidence that larger banks are indeed safer. In addition to U.S. experience in the 19th and early 20th centuries, failure rates in Australia, Canada, Scotland, and England did decrease substantially once banks had consolidated (see references in Section 7). Thus, it seems reasonable to assume that diversification through size does add value over a significant range.

Intermediaries Optimally Issue Debt-Like Contracts to Investors. This follows under a variety of assumptions. If intermediary returns are costly to verify, debt is optimal because it minimizes the situations in which investors must verify the intermediary's true return. If the intermediary's effort invested in monitoring or evaluating risks is unobservable, debt is optimal among a wide class of contracts as a means of inducing effort (Innes, 1990). Finally, debt is the least risky limited-liability contract, which may make it attractive to risk-averse investors. Empirically, many intermediaries are highly leveraged: in the U.S., banks, finance companies, and life insurers all have average leverage ratios of roughly 90% or more (see the Federal Reserve's *Flow of Funds*).

Investors Find It Very Costly to Coordinate Their Actions. This seems especially reasonable for intermediaries with large numbers of small individual investors. Even institutional investors find coordinated action difficult when they are present in numbers; for example, most public corporate debt is held by institutional investors, yet debt renegotiations outside bankruptcy are rarer when more debt is public, hence diffusely held (Gilson *et al.*, 1990).

B. *Specific Assumptions*

Although the features just listed can be found in many models of intermediation, for purposes of exposition I use delegated monitoring with costly state verification. What follows is a streamlined form of this model that incorporates imperfect diversification.²

A continuum of identical risk-neutral investors are endowed with a generic good; the total sum of their endowments is N . Investors can consume

² The seminal work in this area of Diamond (1984) and Williamson (1986) assumes that perfect diversification is possible. The more realistic case of imperfect diversification has been explored by Winton (1995) and Yanelle (1997), who analyze finite environments, and Krasa and Villamil (1992), who analyze the case where project risk has a systematic component.

their endowment or invest it elsewhere. There is also a continuum of entrepreneurs endowed with projects that require funding; each entrepreneur's project requires the endowments of K investors and returns a random amount of generic consumption good at the end of the period. For simplicity, assume that the total mass of entrepreneurs greatly exceeds N/K ; then investors have relative market power and entrepreneurs only receive their reservation return.³

Investors can only observe ("monitor") a project's return at effort cost C . Since an unmonitored entrepreneur has incentive to misrepresent returns (so as to pay investors less and pocket any difference), optimal investment contracts specify the reported returns and associated payments that lead to monitoring so as to minimize investors' expected monitoring costs. Williamson (1986) shows that the optimal symmetric investment contract is debt, but this leads to duplication of monitoring efforts. Let $\underline{v}$ be the maximum of investors' utility from either consuming their endowment or directly investing in entrepreneurs' projects. For simplicity, I abstract from how this "autarky" return depends on the distribution of project returns. Alternatively, one can assume consumption dominates direct investment.

As an alternative, investors can invest in an intermediary ("bank"), which gives the investors debt contracts and invests their endowments in the projects of many entrepreneurs. Banks are run by risk-neutral agents with a reservation return of zero. Since the bank is the only monitor of each project it invests in, the per-unit cost of monitoring each entrepreneur is reduced to C/K . Of course, investors must now monitor the bank, but if project returns are imperfectly correlated, the per-unit risk of monitoring a sufficiently diversified bank can be reduced significantly. I assume that a sufficiently diversified bank can indeed offer investors a return in excess of $\underline{v}$.

Within a period, the sequence of events is as follows. First, banks publicly and simultaneously offer stated rates to investors, after which investors simultaneously choose where to invest their funds. Once banks have raised funds, they invest in entrepreneurs' projects as already described. At the end of the period, entrepreneurs report their returns, which banks monitor as specified by contract. After receiving payments from entrepreneurs, banks report their portfolio returns to investors; if the per-unit return is

³ Winton (1995) considers a setting where the supply of entrepreneurs is endogenous, and banks must compete for borrowers as well as investors. Because borrowers' probabilities of getting loans depend on how many deposits their bank has raised, *borrower* beliefs must be modeled, and investors and borrowers form beliefs based on banks' lending *and* deposit rates. Nevertheless, the current paper's thrust would be unchanged: coordination is even more important under two-sided competition, so coordination problems and fragmentation would be even more likely in a young banking system, and borrower and investor beliefs would be even more likely to give incumbents an advantage in a mature system.

less than the bank's stated rate D , the bank's investors monitor. Payments are made to investors as per contract, all agents consume, and the period ends.

Denote the asset portfolio return of a bank of size n (i.e., a bank with mass n of investors) as $\tilde{X}_n$. If $K\mu$ is the expected project cash flow net of an entrepreneur's reservation return, and the bank funds n/K entrepreneurs, then $E[\tilde{X}_n]$ is $n\mu$.⁴ The precise distribution $G_n(x)$ of $\tilde{X}_n$ depends on the correlation between entrepreneurs' projects, which I abstract from. However, so long as different projects are not perfectly correlated, the risk of the per-unit portfolio return $\tilde{X}_n/n$ decreases in the sense of Rothschild and Stiglitz (1970) as n increases; that is, bigger banks are better diversified.⁵ The distribution of the per unit return $H_n(\cdot)$ is defined by $H_n(x/n) = G_n(x)$. Finally, let c be the bank's expected cost of monitoring per unit invested, i.e., c equals C/K times the probability with which the bank monitors each entrepreneur.

If the bank issues each investor debt with "stated rate" (principal plus interest) D , each investor receives cash payments equal to $\min\{D, \tilde{X}_n/n\}$, and must monitor the bank whenever $\tilde{X}_n/n$ is less than D . Letting $r(n, D)$ be an investor's *expected* cash payment, and integrating by parts, an investor's expected return $v(n, D)$ is

$$v(n, D) = r(n, D) - C \cdot H_n(D) = D - \int_0^D H_n(x) dx - C \cdot H_n(D). \quad (2.1)$$

Similarly, the bank's expected profits $\pi(n, D)$ are given by

$$\pi(n, D) = n \cdot [\mu - c - r(n, D)], \quad (2.2)$$

and the bank's average profit margin $m(n, D)$ is

$$m(n, D) = \frac{\pi(n, D)}{n} = \mu - c - r(n, D). \quad (2.3)$$

The next lemma establishes some useful features of investor returns and bank profits.

⁴ I am abstracting from credit rationing, which occurs if lending at a rate that gives entrepreneurs their reservation return is dominated by lending at a lower rate and reducing verification costs. See Williamson (1986) and Winton (1995) for analysis of intermediation's effect on credit rationing.

⁵ While some might argue that most of the gains from diversification of independent risks are achieved with relatively small numbers of different projects, there may be sectoral and systematic components leading to significant cross-correlations between groups of projects. This could be explicitly modeled by assuming that groups of entrepreneurs of some measure ε form communities or sectors, and then specifying a correlation structure across such groups.

LEMMA 2.1 (Effects of Bank Size and Stated Rate on Investor Returns and Bank Profits). (i) *All else equal, an increase in a bank's stated rate D increases an investor's expected payment $r(n, D)$, increases an investor's expected monitoring costs, and decreases expected bank profits $\pi(n, D)$ and profit margin $m(n, D)$.*

(ii) *All else equal, an increase in bank size n increases an investor's expected payment $r(n, D)$ and decreases expected bank profit margin $m(n, D)$. If bank asset returns are normally distributed, an investor's expected monitoring cost decreases if the stated rate D is less than the mean per-unit asset return μ , and increases if D is greater than μ .*

The proof of this and other results is in the Appendix. The effects of a change in stated rate are intuitive: the expected payment to investors rises, hurting bank profits and margins, but the bank is more likely to default, increasing investors' expected monitoring costs. The effects of a change in bank size follow from the impact of improved diversification. Since debt payments are a concave function of bank asset returns, reducing the per-investor risk of asset returns increases investors' expected returns: the density of returns "tightens" about its mean μ , reducing expected shortfalls on the debt (Rothschild and Stiglitz, 1970). This in turn hurts the bank's expected profit margin. However, although the expected payment to each investor rises, the probability of default need not decrease: indeed, if the stated rate D is above the mean return μ , tightening the density about μ may make default more likely.

In much of what follows, I impose two simplifying assumptions on the behavior of expected investor monitoring costs.

Assumption 1. Suppose a bank of size n offers stated rate D such that investor expected return $v(n, D)$ exceeds $\underline{v}$ and expected bank profits $\pi(n, D)$ are nonnegative. Then the bank's chance of default declines with size (i.e., $\partial H_n(D)/\partial n < 0$).

This implies that larger banks always *potentially* dominate smaller banks.⁶ This assumption seems generally plausible in the context of delegated monitoring. For example, if bank returns are roughly normally distributed, Lemma 2.1(ii) implies that the bank's chance of default increases with diversification only if the stated rate D exceeds the mean return μ ; this implies that the bank defaults more than half the time, which seems unlikely to be a general characteristic of viable banks. Nevertheless, small banks

⁶ To see this, consider two banks, one of size n , the other of size $n' > n$. By Lemma 2.1(ii), $r(n, D) < r(n', D)$, and, so long as $v(n, D) \geq \underline{v}$, $C \cdot H_n(D) > C \cdot H_{n'}(D)$. By lowering its rate to a level D' at which $r(n', D') = r(n, D)$, the large bank reduces investor monitoring costs further; thus investors are better off, while the large bank's profit margin is the same as the smaller bank's.

may have to offer rates in excess of μ , and there may be other unmodeled diseconomies of scale, so I consider the effects of weakening this assumption at the end of Section 3.

Assumption 2. Suppose a bank of size n offers stated rate D such that expected bank profits $\pi(n, D)$ are nonnegative. Then investors' expected return $v(n, D)$ rises with D (i.e., $\partial v(n, D)/\partial D > 0$).

Assumption 2 says that, if the stated rate D is low enough to allow a bank of size n to earn positive profits, increasing D increases an investor's expected payment $r(n, D)$ more than it increases the investor's expected monitoring costs $C \cdot H_n(D)$. Suppose the distribution of bank asset returns has the monotone hazard rate property, as do many common distributions. Then Assumption 2 holds so long as monitoring costs C are not too high; otherwise, investor returns $v(n, D)$ are first increasing, then decreasing, in D .⁷ As discussed at the end of Section 3, the main effect of weakening this assumption is to weaken rate competition among banks.

For any bank size n , define $D_0(n)$ as the stated rate that sets expected bank profits equal to zero, so that $r(n, D_0(n)) = \mu - c$. Since $r(n, 0)$ is 0, $r(n, \infty)$ is μ , and $r(\cdot)$ increases in D , $D_0(n)$ always exists and is unique; also, since $r(\cdot)$ increases in n , $D_0(n)$ is decreasing in n . Similarly, for any n , define $\underline{D}(n)$ as the lowest stated rate D for which investors' expected return $v(n, D)$ equals $\underline{v}$. Assumption 1 (chance of default falls with size) guarantees that $\underline{D}(n)$ is decreasing in n with steeper downward slope than $D_0(n)$.⁸

Since I assume that banking is viable and larger banks can dominate smaller ones, $\underline{D}(N)$ must be less than $D_0(N)$. It seems reasonable to suppose that very small banks are not sufficiently diversified to be preferable to direct lending, so that $\underline{D}(0)$ is strictly greater than $D_0(0)$; this assumption is not essential to my results, but it simplifies analysis. Under this assumption, define $\underline{n}$ as the unique size for which $\underline{D}(n)$ and $D_0(n)$ are equal; $\underline{n}$ is the minimum size a bank must achieve to offer both itself and its investors returns better than autarky.

Figure 1 shows a numerical example of $\underline{D}(\cdot)$ and $D_0(\cdot)$. In this example, total market size N is 10. Bank asset returns are normally distributed with mean $\mu = 1.1$ and variance $0.0008 + (0.0012/n)$, so some risk is undiversifiable (a bank of size 10 has returns with 3% standard deviation). Investors have "autarkic" return 1.03, their cost of monitoring the bank C

⁷ $\partial v(n, D)/\partial D = 1 - H_n(D) - C \cdot h_n(D)$. If the hazard rate $h_n(D)/[1 - H_n(D)]$ increases in D , $\partial v(n, D)/\partial D > 0$ for low D and switches sign at most once. The uniform, normal, and gamma distributions and their truncated versions all have this property (Barlow and Proschan, 1975).

⁸ Winton (1990) shows that $\underline{D}(n)$ has these same basic properties in a model where intermediaries provide insurance, so that $v(n, D) = E[u(\min\{D, \bar{X}_n/n\})]$ for some concave u .

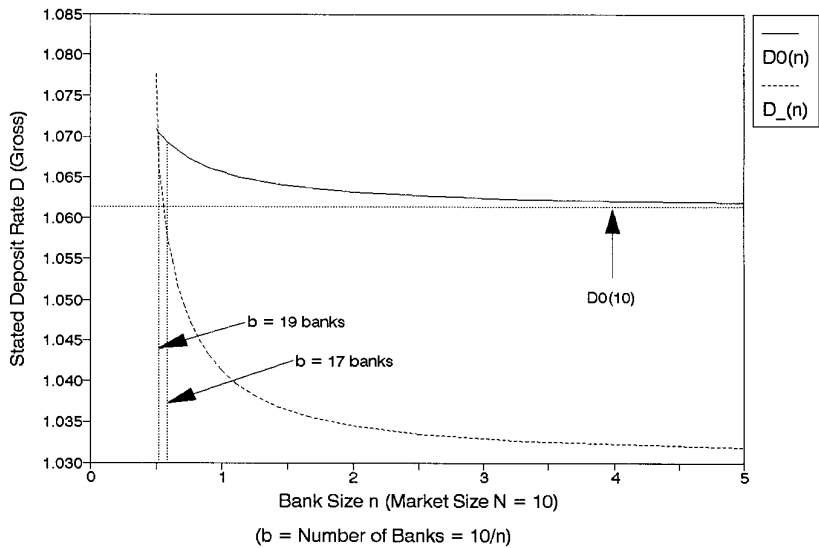


FIG. 1. Bank and investor reservation rates.

is .10, and the bank’s cost per unit of assets c is 0.04. In this example, minimum viable bank size $\underline{n}$ is approximately 5.2% of total market size.

3. RATE COMPETITION AMONG BANKS IN A
STATIC SETTING

This section examines competition in a single-period setting, which establishes the key themes of the paper and serves as a building block for the multiperiod analysis in Section 4. After examining how investors react to bank rates, I look at bank competition in two settings: a relatively new system where no banks are well established, and a mature system where some banks are well established. As argued below, these two settings have very different implications for which investor beliefs are plausible, affecting how investors react to bank rates and thus how banks compete with one another. At the end of the section I discuss some alternative investor beliefs, especially with regard to weakening Assumptions 1 and 2 from the previous section. Throughout, my analysis assumes that equilibrium must be subgame perfect: starting from any possible sequence of events through a point in time, the subsequent behavior of banks and investors satisfies “best response” (is a Nash equilibrium).

A. Investor Behavior

For the moment, suppose the number of banks is fixed at b , either through regulatory fiat or historical accident; later, b will be endogenized under free entry. Solving for equilibrium through backward induction, the first question is how investors behave when faced with rates $\{D_i\}_{i=1}^b$, where D_i is bank i 's stated rate. Subgame perfection requires that, for *any* such set of rates, each investor chooses a bank that maximizes her expected return. This expected return depends not only on her bank's rate, which she can observe *ex ante*, but on the bank's size, which depends on how many investors choose to use that bank. Thus, an investor's actions depend on the relationship between the stated rates banks offer and the share of the market each bank is expected to attract. Of course, in equilibrium, investors' beliefs must be consistent with their eventual behavior; i.e., if a bank is expected to have a certain market share, this must be the bank's share in equilibrium.

Let $n_i(D_1, \dots, D_b)$ be investors' expectation of bank i 's size, given rates $(D_1, \dots, D_b)$. From the preceding discussion, a set of beliefs $\{n_1, \dots, n_b\}$ is consistent with equilibrium if

$$v(n_i(\cdot), D_i) = v(n_j(\cdot), D_j) \geq v(n_k(\cdot), D_k) \text{ for all } i, j, k \text{ such that} \quad (3.1)$$

$$n_i(\cdot) > 0, n_j(\cdot) > 0, \text{ and } n_k(\cdot) = 0,$$

$$v(n_i(\cdot), D_i) \geq \underline{v} \text{ for all } i \text{ such that } n_i(\cdot) \geq 0, \text{ and} \quad (3.2)$$

$$\sum_{i=1}^b n_i(\cdot) \leq N, \text{ with equality whenever } v_i(n_i(\cdot), D_i) > \underline{v} \text{ for some } i. \quad (3.3)$$

Condition (3.1) says that individual investors must be indifferent among banks with positive market shares (otherwise, they would not be maximizing their expected returns), but they may strictly dislike banks that no one uses; (3.2) requires that, if a bank has positive market share, investors prefer the bank over autarky; finally, (3.3) assures that beliefs are consistent with market clearing.

A given set of bank rates may permit several sets of beliefs that are consistent with equilibrium. For example, as long as bank i 's rate D_i is not too much lower than other bank rates, setting i 's size to N and all other bank sizes to 0 is an equilibrium; this could be permuted, or several banks could have nonzero sizes, etc. However, in the current setting, some of these beliefs are more plausible than others.

Consider rates $\{D_i\}_{i=1}^b$ and an associated set of beliefs about bank sizes $\{n_i\}_{i=1}^b$. Suppose bank 1 increased its rate from D_1 to $D_1 + \varepsilon$. So long as $v(n_1, D)$ is increasing in D for D between D_1 and $D_1 + \varepsilon$, any investors

planning to use bank 1 would now be strictly better off, so they should still plan to use bank 1—and investors on the margin at other banks would now prefer to go to bank 1 as well. Indeed, by Assumption 1, $v(n, D)$ increases in n , so the resulting increase in bank 1's size would make investors there even better off. Thus, it seems reasonable that expected bank size $n_i(\cdot)$ should be weakly increasing in bank i 's own rate D_i so long as $\partial v(n_i, D)/\partial D$ is positive at $D = D_i$.

By Assumption 2, this condition holds whenever a bank earns positive profits. It follows that the only way for a profitable bank to increase its market share is to raise its rate. However, this does not necessarily imply that the bank that posts the highest stated rate has the largest market share. If one bank is expected to be larger, hence more diversified, then its stated rate can be lower than those of other banks and still give investors a higher expected return; indeed, I analyze beliefs of this sort in Section 3.C below. Instead, the assumption that beliefs are “increasing” simply says that a profitable bank does not hurt its market share by raising its rate.

This leaves open the question of whether investors think that, all else equal, some banks will be bigger and better diversified. The answer depends on the nature of the banking market. In a new banking system, or one in which the market is growing so fast that banks' current investor bases are small relative to the volume of new investors, investors have no basis for believing that banks are different from one another. Since coordination among investors is prohibitively costly, the only plausible beliefs are that any two banks offering the same rate will have the same market share. By contrast, if some banks are well established, it seems reasonable that investors would think that these incumbent banks are likely to have larger share at any given rate than a new bank would; i.e., past market shares are a guide to future sizes.

In either case, given how investors react to the bank rates that they see, banks set their rates in a Nash equilibrium: i.e., each bank i must choose its rate D_i to maximize its profits $\pi(n_i(D_1, \dots, D_b), D_i)$, taking the other banks' rates as given. I now explore each setting in turn.

B. *Bank Competition in a Young Banking System*

Suppose that the banking system is young and no banks are well established. As just noted, coordination problems suggest that any two banks with the same stated rate will have the same market share. Although it is always possible for investors to have the self-fulfilling belief that, regardless of rate, no bank is viable, for simplicity, I assume beliefs are as “optimistic” as possible.⁹ Specifically, suppose b' banks have the highest rate D , and

⁹ Autarky is always an equilibrium because, to be viable, a bank's size must be at least $\underline{n}$: if investors believe that no bank will have investors, then no bank can profitably offer a rate that tempts any one investor to come, making these beliefs self-fulfilling. See also footnote 19 in Section 6.

that, if these b' banks split the market, investors' expected return exceeds $\underline{v}$ (so $D \geq \underline{D}(N/b')$); then I assume that these banks do indeed split the market ($n_i(\cdot) = N/b'$ for these banks). Otherwise, if D is less than $\underline{D}(N/b')$, investors do not use banks ($n_i(\cdot) \equiv 0$).

Now consider a candidate equilibrium in which all b banks offer the same rate D . If banks are to be viable, D must weakly exceed $\underline{D}(N/b)$. D must also be no more than $D_0(N/b)$; otherwise, bank profits would be negative, and any one bank would benefit by lowering its rate and dropping out of competition. Notice that, given the absence of search or switching costs (see Section 5), a bank that raises its stated rate captures the whole market: increasing its rate makes it more attractive to investors even if its share does not increase, and, as more investors choose the bank, the resulting increase in n increases their expected return even further, while the drop in their market share hurts expected returns offered by competitors. Thus, each bank must compare the profits from increasing its rate and thus its market share with its candidate equilibrium profits $\pi(N/b, D)$.

From Lemma 2.1, bank profits $\pi(n, D)$ decrease in D , and the bank's profit margin $m(n, D)$ decreases in n . Also, $\partial\pi/\partial n$ equals $m(n, D) - n \cdot \partial r/\partial n$, so, all else equal, if its profit margin is not too large, a bank's expected profits decline with an increase in size. Intuitively, diversification's negative impact on the bank's overall profit margin (the debt overhang effect) outweighs the advantage of earning profits on the marginal investor. However, by assumption, if the bank's profits are nonnegative, investors' expected return is increasing in D , and the only way to increase share is to raise D , which is unprofitable. This leads to the next result.

PROPOSITION 3.1 (Equilibrium in a New Banking System with b Active Banks). (i) *Let $\underline{b}$ be the greatest integer less than or equal to $N/\underline{n}$. Then in a new banking system with $b \leq \underline{b}$ banks, there are equilibria in which the banks all offer the same rate D and share the market equally. No equilibrium exists with more than $\underline{b}$ banks having positive market shares.*

(ii) *In these equilibria, the banks' stated rate D can be anywhere in the interval $[D^*(N/b), D_0(N/b)]$, where $D^*(n)$ is the smallest rate such that (a) $D^* \geq \underline{D}(n)$ and (b) $\pi(n, D) \geq \pi(N, D)$ for all $D \geq D^*$. It follows that $\underline{D}(N/b) \leq D^*(N/b) \leq \max \{D_0(N), \underline{D}(N/b)\}$.*

(iii) *If $\partial^2 r(n, D)/\partial n \partial D > 0$, (ii) describes all equilibrium rates. One case where this condition holds is when $\tilde{X}_n/n$ is normally distributed and the rate D is below the mean per-unit asset return μ .¹⁰*

¹⁰ More generally, $\partial r(n, D)/\partial D$ equals $1 - H_n(D)$; this increases with n if $H_n(D)$ decreases in n .

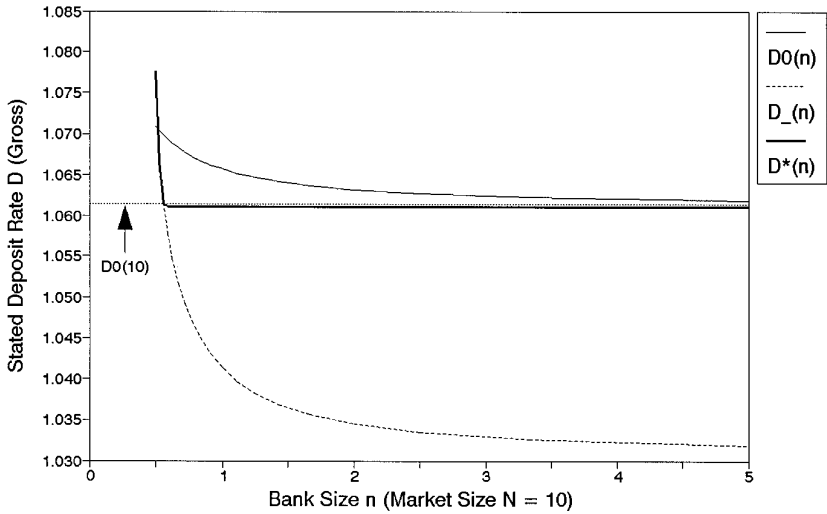


FIG. 2. Minimum static collusive rates $D^*(n)$.

Thus, whenever b banks can viably share the market, there are equilibria in which they do so. If the condition in part (iii) holds, the most profitable of these equilibria has a bank rate of $D^*(N/b)$; this condition requires that the impact of an increase in a bank's stated rate on an investor's expected payments is greater the more diversified is the bank. Figure 2 graphs the relationship between D^* and n for the numerical example already given.

Here and throughout the paper, I require that any collusion among banks must be consistent with subgame perfection; that is, collusion (tacit or explicit) must be self-enforcing. Since there are many subgame perfect equilibria with b active banks, a group of b banks may collude by picking the most profitable of these equilibria, i.e., a rate $D^*(N/b)$. Indeed, unless banks merge or can credibly make side payments to compensate some banks for being less active, this is the best b banks can do in this static setting. In a symmetric equilibrium, industry-wide profits equal profit margin times the total market's size N , so the profit margin $m(N/b, D^*(N/b))$ serves to measure the static potential for collusion.¹¹

¹¹ Requiring that collusion be self-enforcing follows the industrial organization literature (see Tirole, 1988) and seems consistent with the historical examples of bank collusion discussed in Section 7. The use of merger to increase collusion and market power is discussed in Section 5. Contractual arrangements with side payments among banks would require costly ex post verification and might well be illegal.

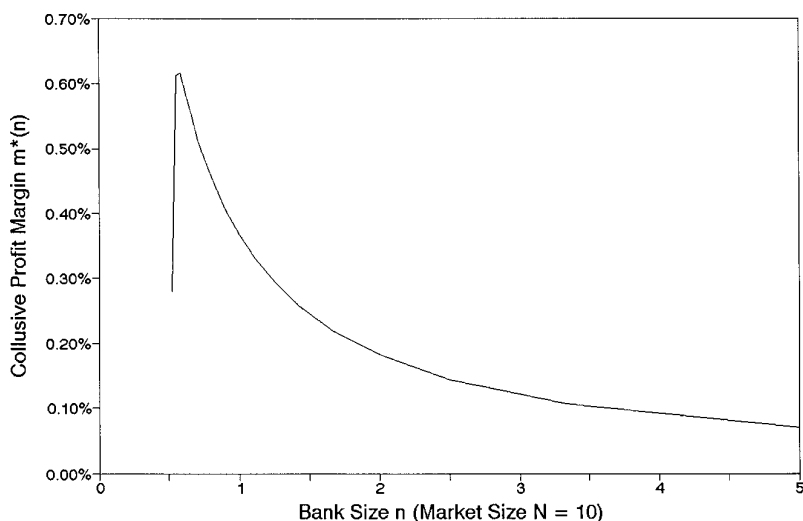


FIG. 3. Maximum static collusive margins $m^*(n)$.

PROPOSITION 3.2. Define the (static) collusive profit margin as $m^*(n) = m(n, D^*(n))$. (i) If $D^*(n) = \underline{D}(n)$, $m^*(n)$ increases with n . (ii) If $D^*(n) > \underline{D}(n)$, $m^*(n)$ decreases with n if $\partial \pi(n, D^*(n)) / \partial n \leq 0$.

When $D^*(n) = \underline{D}(n)$, banks can capture investor surplus in its entirety; by the assumptions of the model, this surplus increases with bank size. This case is likely to hold when the number of banks is close to the maximum viable number $\underline{b}$: here, diversification effects are strongest, so profits are most likely to decrease with bank size, reducing banks' incentive to compete for market share. As the number of banks falls below $\underline{b}$, banks are larger, reducing the marginal impact of diversification and thus the debt overhang effect. Eventually, debt overhang effects are small enough that $D^*(n)$ is above $\underline{D}(n)$; i.e., rates must be higher and margins smaller before debt overhang offsets the benefits to increasing market share.

Thus, although there is generally a range of bank sizes bounded below by $\underline{n}$ for which collusive margins increase with bank size (and thus decrease with the number of banks), once banks are sufficiently large, collusive margins may begin to decrease with size. Figure 3 illustrates this for the numerical example: here, $m^*(n)$ peaks at $n = 0.59$, or $b = 17$ banks. However, as shown in the next section, if banks interact repeatedly, smaller numbers of banks may be able to do better than $m^*(n)$.

Until now, the analysis has assumed that the number of banks is fixed exogenously. When banks are not yet well established, this assumption

makes sense only if regulators explicitly fix the number of bank charters granted. The next result clarifies which equilibria remain when free entry is allowed and the number of banks is determined endogenously.

COROLLARY 3.3 (Equilibrium in a New Banking System with Free Entry). *Suppose any number of banks can freely enter the market and offer rates to investors. An equilibrium with $b \leq \underline{b}$ banks is possible if and only if $D_0(N) \leq \underline{D}(N/(b+1))$; if this holds, the equilibrium rate D can be anywhere in the interval between $\max \{D_0(N), \underline{D}(N/b)\}$ and $\min \{D_0(N/b), \underline{D}(N/(b+1))\}$. This interval always exists for the maximum number $\underline{b}$ of viable banks.*

The intuition here is that entry is always profitable if one of two conditions holds: either banks post rates that make splitting the market with one more bank both profitable and feasible, or banks post rates such that capturing the market with a slightly higher rate yields positive profits. A free-entry equilibrium with b banks is possible if and only if neither condition holds, but one or both hold if $D_0(N)$ exceeds $\underline{D}(N/(b+1))$, which is likely for few banks (small b). As b increases, $\underline{D}(N/(b+1))$ increases, so it is more likely that an interval of free entry equilibria exists; this becomes a certainty when b reaches the maximum number of viable banks $\underline{b}$. Thus, under symmetric beliefs, free entry tends to lead to a more fragmented banking system.

This can be seen in Fig. 1. Here, $D_0(N)$ exceeds $\underline{D}(n)$ when n is 0.59 or greater (b is 17 or less); thus, no equilibrium with 16 or fewer banks is possible when entry is allowed. However, free-entry equilibria do exist with 17 or more banks.

Free entry complicates collusion among banks in several ways. First, Corollary 3.3 clearly shows that the potential for collusion is diminished by free entry (even if b banks wanted to collude at a rate below that permitted in Corollary 3.3, they could not possibly bribe all potential entrants to stay away). Also, since the number of active banks is now endogenous, banks that want to collude may not know who all their potential “partners” are, making it difficult to coordinate on the correct equilibrium rate, and large numbers of potential entrants make it difficult for banks to coordinate on the number of banks b that leads to most the profitable equilibrium in Corollary 3.3.

To summarize, when the banking system is young and investor beliefs have not settled on a group of incumbent banks, banks have some ability to earn “collusive” profits if entry is limited. If free entry is allowed, there is a tendency for the banking system to fragment and for collusion to break down, reflecting both investors’ and banks’ inability to coordinate their beliefs and actions.

C. *Bank Competition in a Mature Banking System*

Suppose now that banks have existed for some time and market growth has slowed. In such a setting, it seems plausible that, *all else equal*, investors will use a bank's past market share as an indication of other investors' likelihood of using that bank at present: since larger banks are better for investors, investors want to coordinate their choices among multiple banks with the same stated rate, and past market share is a cheap coordinating device. As I now show, this creates incumbency power.

First, suppose a group of banks has dominated the market recently. Investors believe these "incumbent" banks will continue to dominate unless an entrant offers a rate that would make investors better off if the entrant got the same market share that incumbents would otherwise have. Effectively, as long as they offer the highest rate, incumbent banks continue to dominate the market, splitting it symmetrically; however, if an entrant offers the highest rate, it will dominate. These beliefs give minimal preference to incumbents over entrants, which seems consistent with a setting where the incumbents have only lately arrived at dominance and investors are still on the lookout for better alternatives.

COROLLARY 3.4 (Entry When Incumbents Have Just Achieved Dominance). *Suppose a group of b incumbents have recently gained dominance in the market (so $b \leq \underline{b}$), but additional banks can freely enter and offer rates to investors. Under the beliefs just described, an equilibrium with the b incumbent banks splitting the market is always possible; the equilibrium rate D can be anywhere in the interval between $\max \{D_0(N), \underline{D}(N/b)\}$ and $D_0(N/b)$, which always exists.*

Corollary 3.4 shows that even this slight advantage of incumbency allows incumbents to maintain market dominance, preventing fragmentation. In a new market, investors have no way to coordinate their choice of institutions, so an entrant can further split the market by offering the prevailing rate; the reduction in bank size increases bank risk and increases expected profit margins, giving the entrant positive profits. By contrast, under beliefs with the slight incumbency advantage described here, investors coordinate on the incumbents unless an entrant offers a higher rate, ruling out the entry strategy of profiting through further fragmentation.

The longer the incumbent banks have dominated the market, the more skepticism investors may display about the viability of entrants. One way in which this should manifest itself is that, in order to attract any investors at all, an entrant must post a rate D that offers investors an attractive expected return even if the entrant's size n is *smaller* than those that the

incumbents would otherwise have.¹² Not surprisingly, such beliefs further reduce entry's potential to restrict collusion.

At an extreme, suppose that the incumbents have dominated the market so long that no investor will go to an entrant unless she would be better off even as the entrant's only customer. Thus, the entrant's rate must be at least $\underline{D}(0)$ in order to attract investors, which leads to the next result.

COROLLARY 3.5 (Entry When Incumbents Are Well Established). *Suppose a group of b ($\leq \underline{b}$) banks has dominated the market for a long time, but additional banks can freely enter and offer rates to investors.*

(i) *Under the beliefs just described, an equilibrium with the b incumbent banks splitting the market is always possible.*

(ii) *The set of possible equilibrium rates is the same as given in Proposition 3.1, so that the collusive profit margin is as given in Corollary 3.2.*

Since the minimum viable bank size $\underline{n}$ exceeds zero, a bank offering a rate of $\underline{D}(0)$ or more loses money if its size remains zero, and any increase in scale simply increases its expected losses at this rate.¹³ Thus, entry is ruled out, and the b incumbents can achieve the collusive profit margin $m^*(N/b)$.

Note that any incumbency advantage makes it easier for banks to collude, since at least there is now a set of banks that are obvious partners. Although the number of incumbents may not be that which maximizes total industry profits, the difficulties in merging discussed in Section 5 suggest that, in the short run, the number of incumbents will be determined by historical accident.

In summary, if investors use past market shares as a coordinating device, incumbents have an advantage over new entrants; to the extent investor beliefs become more "focused" on the incumbents over time, this advantage increases. In the limit, the threat of entry evaporates.¹⁴

¹² This can be viewed as a reduced form of a search model: early on, with fluid numbers of banks, more investors find it worthwhile to incur (small) search costs and check out the competition. As time goes on and the same group of banks are effectively the only players in town, investors find search less attractive, so fewer search and notice entrants' rate offers.

¹³ Even if $\underline{n}$ were zero, so that a bank could offer a rate $\underline{D}(0)$ and earn positive profits, attracting a large share of the market at this rate might be unprofitable, since the increase in size would entail much greater diversification and a substantial drop in the bank's expected profit margin.

¹⁴ For simplicity, I focus on situations where investors assume that all the incumbent banks are roughly equal, but beliefs can also support different incumbent banks splitting the market unequally. In these equilibria, the larger bank offers a lower stated rate and has higher expected profit margin (and thus profits) than its smaller rival; the larger bank's reputation for being bigger gives it an advantage.

D. *Alternative Belief Specifications*

The analysis thus far suggests that, in a young banking system, free entry should cause excessive fragmentation, while in a mature system, free entry may not prevent well established incumbents from earning collusive profits. I have assumed that (i) a bank's market share is always weakly increasing in its stated rate, and (ii) for a given rate D , investors are always better off the bigger the bank; however, as I now discuss, these assumptions can be weakened without changing the qualitative results.

Assuming that a bank's market share is weakly increasing in its stated rate is based on Assumption 2, which says that, if a bank raises its rate and its market share doesn't change, its investors are better off. This might not always be true. First, in the delegated monitoring model, if the stated rate D is high enough, investor's expected returns (net of monitoring costs) may decrease in D . Second, in a more general model, banks of the same size might vary in asset quality; if quality is not publicly observed, investors might view a sufficiently high rate as a sign the bank's assets are of worse quality.¹⁵

Even if sufficiently high rates are unattractive to investors, most of my qualitative results are unaffected. Such beliefs may rule out high-rate, low-profit equilibria, since a bank that lowers its rate may be able to improve investor returns and so profitably increase its size, but this does not rule out the possibility of low-rate, high-profit equilibria. Indeed, so long as the rate at which investor returns start decreasing in D is above $D^*(\cdot)$ as defined in Proposition 3.1, the maximum degree of collusion would be unaffected. Free entry would still fragment a new banking system, since this results from the investors' coordination problem and the debt overhang effect of increased size and diversification. Similarly, in a mature system, investor beliefs should still favor well-established incumbents, restricting entry.

However, it is also possible that an increase in bank size at a given stated rate may not always benefit investors. One case was discussed in Section 2: if the stated rate is sufficiently high, an increase in size and diversification increases the chance of default, so higher expected monitoring costs may offset the increase in expected payments on investors' debt contracts. Also, in a more general model, there may be other diseconomies of scale. For example, as per Cerasi and Daltung (1996), there are probably limits to how many borrowers an individual can monitor, so larger banks must have

¹⁵ On practical grounds, it seems unlikely that lower rates are *always* viewed as better: such beliefs would lead to a "race to the bottom," with each bank competing to offer the lowest viable rate (and earn monopoly profits). There is no empirical evidence of such extreme behavior; indeed, evidence suggests that, to the extent banks use deposit rates to compete for market share, they raise them (for example, see Smith, 1984, on the events leading to the banking panic of 1907).

more monitors, which in turn increases free riding and other agency costs. To the extent marginal benefits to diversification tail off as a bank becomes larger, such diseconomies could eventually offset diversification benefits.

Such organizational diseconomies might prevent a bank from capturing the whole market, either because monitoring quality would be too low to attract investors or because monitoring costs would be too high to be profitable. It would now be easier for large and small banks to coexist. Still, so long as diversification economies dominated over some range, the qualitative thrust of the paper would be unchanged: debt overhang would still undermine banks' desire to expand to the optimal size; entry would still excessively fragment a young banking system, and incumbents in a mature system would gain power by becoming the focus of investor beliefs.

The same is true if the stated rate is so high that an increase in diversification increases the probability of bank default. Indeed, this may reinforce the excessive fragmentation of a young banking system, since the stated rate is most likely to be high when banks are small. If an expansion at current rate levels or higher is unattractive to investors, the many small banks may be completely unable to increase their size through rate competition, even though it is quite possible that this structure is dominated by one with larger banks and lower stated rates.

4. A SIMPLE MODEL OF DYNAMIC COMPETITION AND COLLUSION

Although the static model is suggestive, banks typically exist and compete over extended horizons. Since current market shares may affect future investor beliefs, this may give banks incentive to try to capture larger market shares even at a current loss so as to gain a dominant position in the future. To address this issue, this section presents a simple model of dynamic competition. It turns out that so long as investor beliefs do not assign full dominance (a la Corollary 3.5) to incumbents too quickly, a dynamic setting can actually increase the potential for collusion among banks. Then, in Section 5, I discuss some additional considerations that suggest that the pattern discussed in the static setting—fragmentation in a young banking system, collusion in a mature one—is likely to hold.

Suppose investors and banks live forever and maximize the discounted sum of period returns, with a common discount factor δ between 0 and 1. (This implicitly ignores the fact that failed banks may be unable to resume business, which is discussed in Section 5.) For simplicity, assume that single-period contracts are still used. Equilibria must be subgame perfect.

Since there is no cost of switching between banks in subsequent periods, investors will choose banks on the basis of the expected return each bank

offers in the current period. However, the arguments of Section 3.C suggest that investors will use available information on past market shares to conjecture how likely other investors are to use each bank in the current period.

Following that discussion, suppose that a bank or group of banks have just achieved positive market share, becoming “incumbents.” At first, these banks should gain only the slight advantage used in Corollary 3.4: faced with a rate offer from an “entrant” (a bank that did not have positive share in the previous period), investors go to that bank only if its return dominates that of the incumbents under the assumption that it would have similar share to theirs. However, if the incumbents maintain their dominance over a longer period, their advantage should grow, eventually reaching that used in Corollary 3.5: to gain share, an “entrant” has to offer a rate that is more attractive than the incumbents even if the entrant has only one investor. Although the transition between these two extremes of incumbency advantage might well be gradual, I model it in the following simple way: if a group of banks has dominated for no more than T periods, they have only the slight incumbency advantage of Corollary 3.4; after T periods have passed, this switches to the strong incumbency advantage of Corollary 3.5. Also, for simplicity, I will assume that the best expected return a bank of size zero can offer investors is $\underline{v}$, so that, once a bank or group of banks are well established as incumbents, no entry is possible. These simplifying assumptions ease analysis without affecting the qualitative nature of the results.

Suppose first that only b banks are competing, entry being ruled out either because they have dominated the market for over T periods or by regulatory fiat. If the banks maintain some collusive rate D period after period, each bank’s discounted profits are $(1 - \delta)^{-1} \cdot \pi(N/b, D)$. If collusion is to be self-enforcing, this discounted value must exceed the discounted profits from a bank’s alternative strategy of raising its rate, capturing the market, maintaining dominance in the face of any retaliation, and eventually reaping the benefits of investor beliefs that strongly favor it over all potential rivals. Under this alternative strategy, the bank’s expected discounted profits are just under $\pi(N, D) + \delta \cdot \Pi^I$, where Π^I is the expected discounted value of a bank’s profits when it has just become the sole incumbent.

Suppose the bank does capture the market and maintains dominance for another $T - 1$ periods. Investor beliefs now favor it so strongly that from now on it is immune to entry and can pay the minimum stated rate $\underline{D}(N)$, earning profits $\pi(N, \underline{D}(N)) \equiv \pi_{\max}$ in every period thereafter; by Assumption 1, $\pi_{\max}$ is the largest single-period profit a bank can earn. However, during the $T - 1$ periods before investor beliefs become so focused, other banks may try to capture the market so that *they* can become the incumbents. Given its initial slight incumbency advantage, the bank can always block such entry by offering the same rate as its rivals; however, since rivals are

willing to lose money now in order to earn future monopoly profits, blocking involves a current loss. This leads directly to the following result.

LEMMA 4.1 (The Value of Sole Incumbency). *Suppose that a bank has just become the only incumbent in the previous period. Then, assuming that investor expected returns $v(N, D)$ increase in the stated rate D for a sufficient range above $D_0(N)$, the value of incumbency is*

$$\Pi^I = \frac{\delta^{T-1}}{1 - \delta^T} \cdot \pi_{\max}, \quad (4.1)$$

where $\pi_{\max}$ equals the maximum per period profit $\pi(N, \underline{D}(N))$. Π^I is increasing in the discount factor δ and decreasing in the number of periods T required to become unchallenged monopolist.

The requirement that investors' expected returns $v(N, D)$ increase in D for rates above $D_0(N)$ (the rate at which the incumbent breaks even) guarantees that rivals can pose a credible threat even at rates that lead to current losses. By contrast, if $v(N, D)$ reached a maximum at $D_0(N)$, the incumbent could block its rivals by offering $D_0(N)$ for $T - 1$ more periods, and Π^I would now equal $[\delta^{T-1}/(1 - \delta)] \cdot \pi_{\max}$. As noted in Section 3, once investors' expected returns begin to decrease in the stated rate, it is harder for competition to discipline incumbents, increasing the advantage and value of incumbency.

PROPOSITION 4.2 (Collusion in a Dynamic Setting). (i) *With b well-established incumbent banks, a stated rate of D can be supported as a repeated equilibrium if and only if*

$$\pi(N, D) + \delta \cdot \Pi^I \leq (1 - \delta)^{-1} \cdot \pi(N/b, D), \quad (4.2)$$

or, equivalently,

$$\frac{\delta^T}{1 - \delta^T} \cdot \pi_{\max} \leq [\pi(N/b, D) - \pi(N, D)] + \frac{\delta}{1 - \delta} \cdot \pi(N/b, D). \quad (4.3)$$

(ii) *For any time T until a new incumbent becomes well established, a sufficiently small discount factor δ guarantees that any rate above $D^*(N/b)$ as defined in Proposition 3.1 can be supported as a collusive equilibrium.*

(iii) *If the discount factor δ is greater than $1 - b^{-1} \cdot m(N/b, D)/m(N, D)$, or $D^*(N/b)$ equals $\underline{D}(N/b)$, then a sufficiently long T guarantees that any rate D above $\underline{D}(N/b)$ can be supported as a collusive equilibrium. Otherwise, a sufficiently long T guarantees that any rate above $D^*(N/b)$ can be supported as a collusive equilibrium.*

As (4.3) shows, the would-be monopolist weighs the discounted value of eventually becoming well established as sole incumbent against two factors: the gain or loss in current profits from trying to capture the market (the first term on the right-hand side) and the discounted value of foregone collusive profits (the second term on the right-hand side). If the collusive rate exceeds $D^*(N/b)$, capturing the market involves a current loss, so a low discount rate means that any future considerations of monopoly profits $\pi_{\max}$ are so far removed as to be negligible; the same is true if the time required to become well established is relatively long. Indeed, if the time to become well established is long enough, and the discount rate is not too low, collusive rates below $D^*(N/b)$ may be possible; although capturing the market gives a current profit, the discounted value of foregone collusive profits is higher still, and the monopoly profits from capturing the market are long delayed. This is analogous to the familiar "Folk Theorem" for repeated Bertrand competition. However, here the discount factor must exceed $1 - b^{-1} \cdot m(N/b, D)/m(N, D)$; because profit margins decline with size, this lower bound is smaller than $1 - b^{-1}$, the limit in the standard Bertrand competition model. Once again, the debt overhang effect makes collusion easier to support, especially for larger numbers of (smaller) banks.

The upshot is that, even in a dynamic setting where higher market share today translates into a competitive advantage tomorrow, once some banks have become well established, it is quite possible that they will collude and maintain their current shares rather than compete more aggressively for additional advantage. Indeed, the possibilities for collusion may be enhanced: rates below $D^*(N/b)$ are not equilibria in a single-period setting, but may be supportable here. Collusion of some sort is most likely if the time to become well established T is not too short.

Now consider the case of a young banking system with no well established banks. From Corollary 3.3, any attempt to earn positive profits in the present is likely to attract entry, resulting in a fragmented market. Even if fewer than $\underline{b}$ banks try to set rates so high that they would lose money, they are unlikely to prevent entry: with more banks at the same rate, profit margins will be higher and possibly positive. Thus, as in the static analysis, entry is likely to be maximal.

Maximal entry suggests that bank profits will not be too high, since even the minimum viable stated rate $\underline{D}(n)$ is close to the breakeven point $D_0(n)$. Nevertheless, even though collusion may have little scope, banks may still prefer to keep things as they are rather than compete aggressively to try and capture the entire market. This is because capturing the market is most expensive for a small bank: the initial rate D is likely to exceed $D_0(N)$, so that the bank actually suffers large losses if it expands. Intuitively, the increase in diversification is greatest for small banks, making their "debt overhang" cost of expansion largest as well. Furthermore, if D exceeds the

mean bank asset return μ , then the argument from the end of Section 3 suggests that a small bank may be unable to capture the market even if it is willing to bear the cost: larger size may increase the bank's chance of default, making it less attractive to investors.

5. ADDITIONAL CONSIDERATIONS FOR DYNAMIC COMPETITION

To simplify analysis, I have omitted several complications. I now discuss those that seem especially relevant to dynamic competition: the effect of a bank's default on investors' subsequent beliefs; a less rigid link between bank size and risk; the role of bank capital; investor switching costs; and mergers as an alternative means of expansion. These concerns tend to reinforce the prediction that young banking systems are excessively fragmented, while mature systems are stable and prone to collusion.

The Impact of a Bank's Default on Investor Beliefs. The analysis in Section 4 assumes that, if a bank defaults in one period, it reopens for business in the next period with no harm done. In reality, such a default might well make investors skeptical about the reorganized bank's viability, and the bank's initial owners are likely to have lost much or all of their stake in any future profits the bank might earn. Moving to the other extreme, assume that the owner of a bank that defaults loses any financial interest in the bank, and that the bank itself now becomes an "entrant." These assumptions reduce the effective discount factor: a bank of size n and rate D survives with probability $1 - H_n(D)$, so the current value of profits $\pi(\cdot)$ next period is only $[1 - H_n(D)] \cdot \delta \cdot \pi(\cdot)$. The lower effective discount factor means that future profits weigh less heavily on the bank's current decisions, moving the situation closer to that analyzed in Section 3.¹⁶

A bank's failure also means that there are fewer incumbents, giving survivors greater market power. Thus, even if the bank system is initially fragmented, as time passes and some banks fail, their more fortunate rivals will gain in market share and diversification, becoming safer and thus placing greater weight on future profits when choosing between collusion and deviation.

¹⁶ There are some additional complications for banks' choice between colluding or deviating to become sole incumbent: deviating increases market share, which generally decreases a bank's chance of default, but it also requires a higher rate and so a higher default probability for $T - 1$ periods. Thus, it is ambiguous whether deviating or not deviating has the higher effective discount factor.

Bank Size and Risk. Although my analysis assumes that an increase in bank size leads to a fixed increase in diversification, real banks do have some control over their industry and geographic concentrations. Also, investors may have difficulty observing the bank's asset mix in a clear and timely fashion; public reports on asset concentrations are infrequent and lagged, and typically show only broad asset groupings. To model these points simply, suppose banks can choose high- or low-risk asset strategies, where for any size the low-risk strategy has the same mean return as the high-risk strategy, but is less risky (better diversified) in the sense of Rothschild and Stiglitz (1970). Also, first suppose that strategies can be changed anew each period and that the choice is unobservable.

If banks do not care about the future (their discount rate δ is zero), they will choose the high-risk strategy: for any stated rate, increasing risk decreases expected payments per investor $r(n, D)$, increasing expected bank profits. Expecting such "risk shifting," investors will demand a higher stated rate. Ex post, banks will not profit from risk shifting, but ex ante, they cannot commit to abstain from it.

If banks do care about the future (δ exceeds zero), increasing risk today reduces the odds of surviving to capture future profits, so sufficiently high future profits deter risk shifting, as in Marcus (1984) and Besanko and Thakor (1993). The bank's relative credibility can have a self-fulfilling effect on its decision. If investors believe the bank will choose the high-risk strategy, they will demand a higher stated rate; by reducing the bank's profit margins and increasing its chance of failure, this increases current debt overhang and decreases future profits' relative importance, making the high-risk strategy relatively more attractive to the bank. Conversely, if investors believe the bank will choose the low-risk strategy, they will accept a lower stated rate, making the low-risk strategy more attractive.

An important caveat is that the low-risk expansion strategy involves choosing and monitoring borrowers that differ more by region or sector, so that the bank must have expertise in a number of different areas. Since such expertise takes time to build and manage, it may not be possible for a bank to expand too rapidly without following a more risky strategy: either specialized and undiversified, or diversified but poorly chosen and worse in a first-order stochastic dominance sense.

Risk-shifting incentives, credibility with investors, and difficulties with rapid expansion are most likely to weigh against the low-risk strategy in a new or rapidly growing system. Since banks are likely to be small and risky, they must pay higher stated rates, making the high-risk strategy relatively more attractive. Also, a bank that wishes to diversify well and build a reputation for doing so will have to pay excessively high rates until investors gradually learn the bank's actual risk profile. Finally, small banks have few relationships or assets in place to constrain their choice and thus have both

greater opportunity to choose a high-risk expansion strategy and greater cost to developing the expertise necessary for a low-risk strategy. Conversely, in a mature system, surviving banks are likely to be larger and safer, have well-established reputations and incumbency power, and have more existing business to constrain risk taking, all of which should encourage collusion and a low-risk strategy.

Bank Capital and Leverage. Although I do not model bank equity capital explicitly, my model is consistent with banks having a constant cost of equity capital and keeping the ratio of equity to total assets fixed as they expand. However, equity capital is likely to have increasing cost due to increased agency problems between bank management and outside shareholders (Besanko and Kanatas, 1996); indeed, larger banks are generally more highly levered, suggesting a tradeoff between diversification and leverage. One might argue that, by offering a higher rate, a bank can always expand its market share profitably, the reduction in profit margin caused by better diversification being offset by the increase in margin caused by higher leverage.

However, this argument ignores the impact of higher leverage on investors' expected returns. As shown in Winton (1995), if an increase in bank size is to make investors better off, their gains from diversification must outweigh their losses from increased bank leverage; and even if the change is attractive to investors, lowering the deposit rate is unlikely to attract them in the first place. Thus, the bank may still find such an increase in size unattractive. Conversely, if the expansion is not in investors' interest, they will not respond. Thus, in a young system, coordination problems may still cause the formation of many underdiversified banks that cannot profitably increase market share.¹⁷ As time passes, surviving banks can increase their equity capital through retained earnings rather than costly external finance, and stronger capital reinforces their advantage over entrants.

Switching Costs. Search or switching costs would tend to reinforce this paper's results. Early on, investors would face even more difficulty in coordinating on a few banks; later, unless the economy was growing rapidly and new investors were flooding the market, banks would find it harder to increase market share via rate competition. Indeed, this paper's focus on beliefs and their evolution over time motivates switching costs that are psychological rather than physical.

¹⁷ Even if capital protects debtholders in a fragmented banking system, bank equity will be very risky, which will be costly if shareholders are risk averse or must monitor bank managers. Comparing the concentrated Canadian banking system with the fragmented American system during 1920–1980, Bordo *et al.* (1995b) find evidence of such costs.

Expansion via Merger. Given the costs of rate competition for market share, expansion through merger may well be more attractive. However, if the merged entity is to lower its rate and reap gains from improved diversification, investors must anticipate that it will retain the merged banks' market shares and choose a low-risk strategy. This suggests that mergers are most likely to occur after enough time has elapsed so that incumbent banks have amassed some advantage in terms of investor beliefs. The pace of mergers will also be slower to the extent there are costs of integrating different banks or investors require time to fully digest the permanence of increased market share.

6. REGULATORY ISSUES

Suppose regulators want to maximize social welfare in this setting. After discussing social objectives in somewhat more detail, I analyze the efficacy of two instruments regulators are likely to have at their disposal: entry restrictions and government insurance for investors.¹⁸ By themselves, entry restrictions reduce excessive fragmentation, but also enhance incumbency power and collusion; conversely, government insurance weakens incumbency power, but may lead to excessive entry and fragmentation. However, under some circumstances, a judicious mix of temporary entry restrictions and government insurance can achieve good results.

A. Social Objectives

Fragmentation. Absent diseconomies of scale, the model implies that larger banks always potentially dominate smaller ones: as per footnote 6, being better diversified, larger banks can guarantee investors the same expected payments with lower stated rates, typically reducing monitoring costs as well. All else equal, a single bank should be optimal, and concentration preferred to fragmentation.

Collusion. If they are few in number, banks may collude. In the basic model, this may actually be attractive: the sum of all agents' utilities is maximized by minimizing total monitoring costs (total consumption is constant); since lower stated rates reduce monitoring costs, a single bank with a stated rate of $\underline{D}(N)$ is optimal. This can easily be achieved by offering one banking license and forbidding entry.

¹⁸ If regulators could optimally choose both the number of banks and their prices (rates), the first-best could easily be attained. Such precision seems unlikely in practice: especially when it comes to pricing, regulation is slow to adapt to changes in the precise level of risk, monitoring costs, etc.

However, this stark result ignores many likely complications. Concern for fairness or the populist pressures of a democracy (Roe, 1997) may make regulators place more weight on the welfare of the many investors than that of the fewer banks. Transfers from banks to investors involve dead-weight costs: in my model, regulators must monitor each bank whenever its transfers are below their maximum level. Also, in richer models, collusion or monopoly cause distortions: for example, if the supply of entrepreneurs is endogenous (Winton, 1995) or firm investment scale is not fixed (Rajan, 1992), a monopoly bank's pricing can cause underinvestment, harming welfare; if there is separation between ownership and control, collusion may increase managerial slack at all investors' expense.

Thus, from a social viewpoint, allowing the banking system to evolve naturally may lead to two types of inefficiency: early on, there will be excessive fragmentation and risk; once the market has matured, costly collusion is likely. I now turn to the question of what regulators can do about this.

B. *Entry Restrictions*

Since fragmentation is socially costly, an obvious solution is to restrict the number of entrants into a young banking system, reducing coordination problems and hastening the formation of a few well-diversified banks. The downside is that these banks immediately become incumbents, gaining an advantage over later entrants; furthermore, bureaucratic rigidities may lead to delays in removing the restrictions, and the longer restrictions are in place, the more heavily investor beliefs will focus on the incumbents, increasing their advantage. Thus, entry restrictions hasten both concentration and collusion.

If banks in other jurisdictions are well diversified, regulators can reduce incumbency advantage by allowing entry by these "foreign" banks: facing less debt overhang, such banks have greater incentive to expand at any given rate. However, informational problems in assessing "foreign" banks may reduce the banks' credibility with investors; even if cheap disclosure is possible, cultural distrust or outright xenophobia may interfere. Such issues will play into the hands of special interest groups (e.g., domestic bankers) that wish to prevent entry; a case in point is that of the European Union, many of whose member countries delayed financial integration to let domestic institutions "bulk up."

C. *Government Guarantees or Insurance*

Another way to weaken incumbency advantage is to insure investors against bank default, funding the insurance either with premia from banks or with general tax levies. With full insurance, investors no longer care

about bank size and risk; since they receive their bank's stated rate for sure, they will choose the bank or banks with the highest stated rate, regardless of incumbency status. Thus, insurance makes the banking market more "contestable" in the sense of Baumol *et al.* (1986).¹⁹

Although it is unlikely that regulators can quickly and precisely assess bank risk (see footnote 18), consider this as a benchmark. If investors were fully insured and insurance correctly priced, a bank of size n and rate D would pay a per-dollar premium $\phi(n, D) = D - r(n, D)$, giving it profits of $n \cdot [\mu - c - D]$.²⁰ Since fully insured investors do not care about bank size, an entrant offering the same rate as incumbents could attract some investors, and entry would be attractive unless bank profits were zero; i.e., $D = \mu - c$. Thus, if regulators initially restricted entry, then lifted restrictions and offered full deposit insurance, the threat of entry would force the few incumbent banks to keep rates at $\mu - c$.

In practice, risk-based premia are likely to be rigid and imprecise; for example, those mandated by FDICIA vary much less across risk classes than do credit spreads on bonds of different credit ratings. At the other extreme, suppose investors are fully insured and banks pay a flat premium ϕ per depositor; now bank profits are $\pi(n, D) - n\phi$, and profit margins are $\mu - c - r(n, D) - \phi$. Since investors care only about stated rates, and $r(n, D)$ increases with n , an additional bank can always offer the same rate as others and profitably split the market. Absent entry restrictions, fragmentation will be extreme: unless the premium ϕ is so high that banks cannot be profitable at any size, banks will be as small as possible (in the absence of fixed costs of operation, each bank would have only one borrower). This is a variant of the well-known result that flat rate deposit insurance encourages risk shifting: investors do not care about bank size and diversification, and new banks profit from being smaller and riskier.

Under full flat-rate insurance, temporary entry restrictions will not prevent fragmentation; banks have incentive to enter even if incumbents are earning zero profits. However, if insurance is not full or there are costs and delays in processing investors' claims, investors care somewhat about

¹⁹ Insurance may have other advantages: ruling out "pessimistic" beliefs that lead to autarky (Diamond and Dybvig, 1983), improving the liquidity of deposits (Gorton and Pennacchi, 1990), and delegating bank monitoring from investors to regulators (Emmons, 1993). On the other hand, there may be costs of transferring funds to investors, and, as monitors, regulators are unlikely to be free from agency problems or error. Since my focus is on the effects of government insurance on bank competition, I ignore any net improvement in social welfare from improved liquidity, reduced monitoring costs, etc.

²⁰ To prevent risk shifting, the premium should be collected after banks set lending rates and terms. However, if collected before banks invest, aggregate investment is reduced; if collected afterwards, monitoring costs must be incurred. My discussion abstracts from these issues.

perceived bank size and risk. Nevertheless, it is still easier for entrants to threaten incumbents than in the uninsured case: even if investors think an entrant will be small, size matters less to investors, so they will switch to the entrant for a lower premium over incumbents' rates. In this less extreme case, a policy of restricting entry, allowing incumbents to gain some advantage in investor beliefs, and then removing entry barriers may lead to a system of relatively few banks posting almost competitive rates.

The upshot is that government insurance does reduce incumbency advantage, but, since it is unlikely it can be perfectly priced in a timely fashion, it increases the system's tendency toward fragmentation. This suggests that regulators may prefer a system with less than full insurance, combined with temporary entry restrictions aimed at preventing initial fragmentation. However, if regulators are not totally selfless and regulatory capture is a concern, entry restrictions may be continued indefinitely at the behest of the incumbents, leading to the collusive outcome already discussed.

7. HISTORICAL EVIDENCE

As suggested in the introduction, historical evidence from Australia, Canada, Britain, and the U.S. is consistent with the model's predictions. I now summarize and discuss this evidence.

Australia. In the second half of the 19th century, Australia's relatively unregulated and rapidly growing economy saw the number of "regular" (trading) banks increase from 8 in 1850 to 23 by 1892, with even greater increases in the numbers of building societies and savings and land banks. During the 1891–1893 depression, many building societies and savings and land banks failed, and 13 of the regular banks suspended and underwent "reconstruction." After the crisis, the regular banks began to merge, their number dropping by half through 1920 (Dowd, 1992); they became more conservative, emphasizing "short-term self-liquidating investments in place of longer, more speculative ones" (Dowd, 1972, p.72); and they were now "able to maintain agreements on deposit rates and foreign exchange rates and margins, and there was no longer any effective competition except for advances" (Dowd, 1992, p. 72).

Canada. Although Canada had only 37 banks when separate provinces first united in 1867, by 1874, this had grown to 51 banks (Neufeld, 1972). Figure 4 shows that the number of banks slowly fell to 10 by 1932, much of the consolidation occurring initially through failure, then increasingly through merger. Up to 1920, losses to depositors were similar to those in the much more fragmented U.S. system (Bordo *et al.*, 1995a), but once the Canadian system had consolidated, it became quite stable; there were no

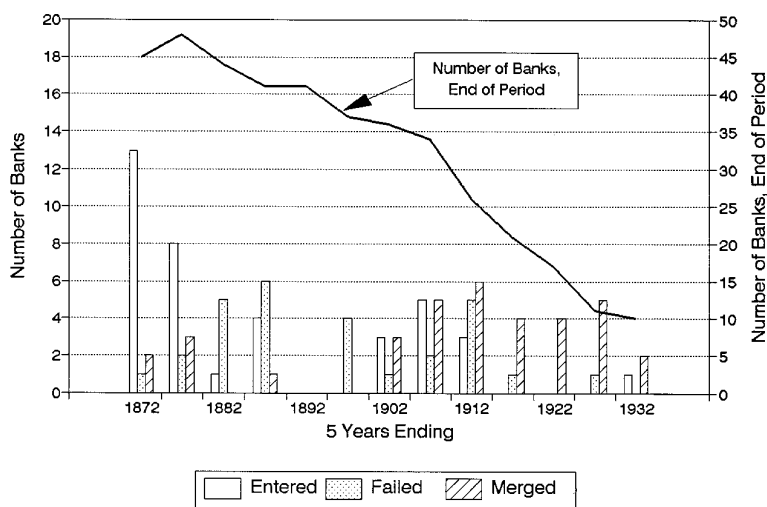


FIG. 4. Entry and exit in Canadian banking, 1868–1932. Source: Neufeld (1972).

failures after 1923, even during the Great Depression. Incumbents were almost immune from entry—from 1912 to 1967, only one new bank formed—and began to collude on loan and deposit rates. This collusive behavior came out in the 1964 Porter Commission report, which helped motivate the introduction of compulsory flat-rate deposit insurance in 1967. After this, “near-banks” (who could join the insurance scheme) became more competitive (Bordo *et al.*, 1995b), and two new banks were formed in 1968–1969 (Neufeld, 1972), suggesting that the incumbents’ advantage was greatly weakened. Indeed, Shaffer (1993) finds little evidence of collusive behavior during the subsequent period.

Britain. Scotland’s well-known free-banking regime had 36 banks in 1825; this fell to 17 by 1850 and 8 by 1913, with merger accounting for most of the consolidations after 1850 (Collins, 1988, pp. 52–53). By 1850, banks were able to support collusion (Checkland, 1975, pp. 391–392 and p. 486).

The banking system in England was initially far more fragmented due to regulatory restrictions, with 650 banks in 1825. Limits on the number of partners per bank were lifted in 1826, and limited liability became generally available in 1862. Consolidation took some time to occur: the number of banks fell to 358 in 1875, and 164 in 1900 (Collins, 1988, p. 52), while the 10 largest banks held 33% of all deposits in 1870, 46% in 1900, and 97% in 1920 (Capie and Rodrik-Bali, 1982, pp. 287–288). Collusion began on some accounts in the 1860s and became total by 1920 (Griffiths, 1973, pp. 6–7).

The banking system was quite stable by the turn of the century (Capie, 1995).²¹

Consistent with the result that mature systems without explicit guarantees should endogenously create large barriers to entry, the British banking system seems to have been virtually immune to entry from the late 19th century until the late 1950s, when “foreign banks and domestic merchant banks, drawing on the growing pool of Euro-dollars, stepped up their intervention in the domestic market place” (Sheppard, 1971, p. 17). When entry did arrive, it came in the form of large money center banks from other countries rather than *de novo* banks, consistent with the notion that well-established, well-diversified rivals would suffer less of a disadvantage in competing for investor funds. In addition, the new entrants raised funds in the Euromarkets, whose institutional investors were probably more sophisticated (and perhaps less culturally biased) than individual investors.

United States. As is well known, regulations in the U.S. produced an extremely fragmented system. However, while this fragmentation is generally viewed as the result of restrictions on branching, deposit insurance seems to have played a role as well. In a study of entry and exit in U.S. banking during the early 20th century, Flood and Kwan (1995) find evidence of excessive entry into banking in general, and find that such “overbanking” was highest in those states that had (flat rate) deposit insurance programs. This is consistent with the notion that fragmented systems are most vulnerable to entry and that deposit insurance further undercuts incumbency advantage.

Discussion. The historical evidence is in many ways consistent with the model’s predictions. In those countries with relatively little regulation (banking in Australia, Canada, and Scotland, insurance in Britain), initial fragmentation and high rates of failure are followed by slow consolidation, first through failure, then increasingly through merger. Even without explicit restrictions, consolidation is linked to increased stability, difficulty of entry, and collusion (Australia, Britain, and Canada). The U.S. experience shows the converse, with a fragmented system inviting continued entry and risk. In an already concentrated system, deposit insurance weakens incumbency power, but need not fragment the market (Canada); by contrast, in a fragmented system, it makes fragmentation even worse (U.S.).

Although organizational diseconomies of scale might also lead to the formation of many small intermediaries, with slow consolidation keeping

²¹ The relatively unregulated evolution of property and casualty insurance in 19th century Britain also displays initial risky fragmentation and slow consolidation (Jenkins, 1984, and Cockerel and Green, 1994). Insurers established a stable cartel by the early 20th century (Sheppard, 1971, pp. 13–17).

step with technological progress, the path of intermediation in the countries just discussed seems to owe as much to the maturity and growth prospects of the intermediary sector as to technological progress per se. This is particularly true of late 19th-century Australia and Canada; for example, the Australian banks were able to operate large networks of over 100 branches (Dowd, 1992), but the economy was in a high growth phase, which would tend to make market shares more fluid and entry easier. The same applies to Canada, where the economy and bank assets as a share of GNP grew rapidly until roughly 1910 (Neufeld, 1972, Chart 3:1, p. 56).

8. CONCLUSION

Existing theories of financial intermediation imply increasing returns to scale linked to diversification. This suggests a simple policy prescription for regulators: the banking sector should be left relatively unrestricted, which should in turn lead to an equilibrium with a few large, well-diversified, and competitive banks. However, the actual pattern of unrestricted banking sector evolution in many countries is at odds with this prediction: high initial entry is followed first by slow consolidation through failure and merger and then by a strong tendency to collusion among survivors.

I develop a model that illuminates this pattern. I show that, in unregulated settings where intermediaries are new or the market is growing rapidly, there should be substantial entry, with many risky intermediaries coexisting: investors cannot coordinate their actions, and debt overhang makes the cost of capturing market share through rate competition highest when the potential for diversification is greatest. Over time, banks will fail, and survivors will gain an incumbency advantage simply by becoming the focus of investor beliefs. Unless intermediaries in other jurisdictions are both well diversified and able to overcome the obstacles to penetrating a foreign market, these incumbents will become relatively immune to entry. This suggests a greater tendency towards collusion in mature banking systems than in other industries with similar concentration: investor beliefs make the market less "contestable" in general, and the debt overhang effect may further reduce competition.

From the regulatory view, these results suggest a tradeoff between fragmentation and collusion. Barriers to entry enhance collusion as well as stability, while government insurance for investors makes the market more contestable but may encourage excessive entry. As a result, regulators may wish to combine less than full insurance with temporary entry restrictions.

The model's results should apply most strongly in situations where gains to additional diversification are large and informational asymmetries both among intermediaries and between intermediaries and investors are espe-

cially costly to resolve. At the present day, these concerns seem to be of decreasing importance in the U.S. and other developed economies, where loan sales and credit derivatives are becoming more efficient and investors are able to obtain more timely and detailed information about intermediaries and their exposures. Nevertheless, to the extent some information asymmetries remain, the role of investor beliefs modeled here will continue to have some relevance, particularly for intermediaries that deal with (relatively unsophisticated) individual investors. Moreover, these concerns should be of critical importance to regulators in developing areas such as Latin America and Eastern Europe, where financial systems are younger and information problems greater.

APPENDIX

Proof of Lemma 2.1. (i) This follows immediately from (2.1) through (2.3) and the fact that expected monitoring costs are $H_n(D)$. For future reference, note that $\partial r/\partial D = 1 - H_n(D)$.

(ii) An increase in n increases bank diversification, reducing the risk of $\tilde{X}_n/n$ in the sense of Rothschild and Stiglitz (1970); it follows from condition (7) in Rothschild and Stiglitz that

$$\frac{\partial}{\partial n} \int_0^D H_n(x) dx < 0, \quad (\text{A.1})$$

which implies that $\partial r/\partial n > 0$. Thus $\partial m/\partial n = -\partial r/\partial n < 0$. Finally, if returns are normally distributed, an increase in n reduces the variance σ_n of $\tilde{X}_n/n$; if D is below μ , $H_n(D)$ falls, while if D is above μ , it rises. Q.E.D.

Proof of Proposition 3.1. Suppose the banks' rate is $D \in [D_0(N), D_0(N/b)]$. If $D < \underline{D}(N/b)$, investors prefer autarky, so such rates can be ignored. Similarly, with more than $\underline{b}$ banks offering the same rate, investors do not believe the banks are viable. Otherwise, each bank receives market share N/b and nonnegative profits. Deviation to a lower rate causes the bank to have zero market share and profits, while deviation to a higher rate leads to a bank of size N with nonpositive profits since its rate is greater than or equal to $D_0(N)$. Since $\pi(\cdot)/n$ is decreasing in n , $\pi(N/b, D_0(N)) > 0$; so long as $D_0(N) \geq \underline{D}(N/b)$, by continuity there is some range of rates below $D_0(N)$ for which it does not pay to raise one's rate and capture the market.

If $D^*(N/b) = \underline{D}(N/b)$, all feasible and profitable rates are equilibria. Otherwise, $\pi(N/b, D^*) = \pi(N, D^*)$. Since $\partial \pi/\partial D = -n \cdot \partial r/\partial D$, so long

as $\partial^2 r / \partial n \partial D > 0$, $\pi(N/b, D) < \pi(N, D)$ for all $D < D^*(N/b)$, so these lower rates cannot be equilibria. Q.E.D.

Proof of Proposition 3.2. (i) If $D^*(n) = \underline{D}(n)$, then the collusive profit margin is $\mu - c - r(n, \underline{D}(n))$. $dm^*(n)/dn = -\partial r / \partial n - (\partial r / \partial D)[d\underline{D}(n)/dn]$. By the Implicit Function Theorem, $d\underline{D}(n)/dn = -(\partial v / \partial n) / (\partial v / \partial D)$; since $\partial v / \partial n = \partial r / \partial n - C \cdot \partial H_n(D) / \partial n > \partial r / \partial n$ ($\partial H_n(D) / \partial n < 0$ by assumption), and $\partial v / \partial D = \partial r / \partial D - C \cdot h_n(D) < \partial r / \partial D$, $d\underline{D}(n)/dn < -(\partial r(n, D) / \partial n) / (\partial r(n, D) / \partial D) < 0$, and $dm^*(n)/dn$ is positive.

(ii) Now, $dm^*/dn = -\partial r / \partial n - (\partial r / \partial D)(dD^*(n)/dn)$. $D^*(n) > \underline{D}(n)$, so $\pi(n, D^*(n)) = \pi(N, D^*(n))$ and $\partial \pi(n, D^*) / \partial D \geq \partial \pi(N, D^*) / \partial D$ (otherwise, $D^* - \varepsilon$ would be an equilibrium rate, contradicting the definition of D^*). Applying the Implicit Function Theorem to the first condition, $dD^*/dn = -[\partial \pi(n, D^*) / \partial n] / [(\partial \pi(n, D^*) / \partial D) - (\partial \pi(N, D^*) / \partial D)]$. Since the denominator is positive, $\partial \pi(n, D^*) / \partial n$ negative implies dD^*/dn positive and $dm^*(n)/dn$ negative. Q.E.D.

Proof of Corollary 3.3. If b banks are earning non-negative profits at some rate D , $\underline{D}(N/b) \leq D \leq D_0(N/b) < D_0(N/(b+1))$; an entrant that also sets its rate to D can profitably and feasibly split the market if and only if $D > \underline{D}(N/(b+1))$. If the entrant sets its rate to $D + \varepsilon$ (where ε is small and positive), it captures the entire market; this is profitable if and only if $D < D_0(N)$. This accounts for the end points of the set of possible equilibria which resist entry; the interval will be empty if and only if $D_0(N) > \underline{D}(N/(b+1))$. Since this interval is contained within the interval specified in Proposition 3.1, these are in fact equilibria. By definition $\underline{b} \leq N/\underline{n}$, but $\underline{b} + 1 > N/\underline{n}$, so $\underline{D}(N/(\underline{b} + 1)) > D_0(N/(\underline{b} + 1)) > D_0(N)$, proving the last part of the corollary. Q.E.D.

Proof of Corollary 3.4. Suppose an entrant posts the same rate as the incumbents; then it gets no share, and profits are zero. The only way to profitably enter is to offer a higher rate and take the entire market, which requires that $D > \underline{D}(N/b)$; such entry will not be profitable if $D > D_0(N)$. Thus, the incumbents can profitably share the market and prevent entry if $D > \max \{D_0(N), \underline{D}(N/b)\}$ and $D < D_0(N/b)$; since this is a subset of the equilibria in Proposition 3.1, it is also a subgame perfect equilibrium among the b incumbents. Q.E.D.

Proof of Corollary 3.5. As outlined in the text, entry cannot be feasible and profitable under these beliefs. Thus, the banks can support any equilibrium from Proposition 3.1. Part (ii) follows immediately. Q.E.D.

Proof of Lemma 4.1. Once the bank becomes well-established as sole incumbent, its discounted profits are $(1 - \delta)^{-1} \cdot \pi_{\max}$. During the $T - 1$ periods preceding this, rivals will be willing to pay up to some rate D_R to

try to get sole incumbency (the value of getting sole incumbency is static, so the rivals' rate D_R is constant), so the incumbent gets $\pi(N, D_R)$ in each of those periods. Thus

$$\Pi^I = \frac{1 - \delta^{T-1}}{1 - \delta} \cdot \pi(N, D_R) + \frac{\delta^{T-1}}{1 - \delta} \cdot \pi_{\max}. \quad (\text{A.2})$$

Any rival would be willing to lose up to $\delta \cdot \Pi^I$ to gain sole incumbency, so $\pi(N, D_R) = -\delta \cdot \Pi^I$. Substituting for $\pi(N, D_R)$ in (A.2) leads to (4.1) in the text. The effects of changes in δ and T are immediate. Q.E.D.

Proof of Proposition 4.2. (i) If a bank captures the market today at a rate just above D , it gets just less than $\pi(N, D)$ this period, and becomes sole incumbent next period, which is worth $\delta \cdot \Pi^I$ today; this equals the left-hand side of (4.2). The right-hand side of (4.2) is the value of the candidate collusive equilibrium. If capturing the market today is not worthwhile, it is not worthwhile later, either (nothing has changed), so this condition is necessary and sufficient. To get (4.3), subtract $\pi(N, D)$ from both sides, rewrite $(1 - \delta)^{-1}$ as $1 + [\delta/(1 - \delta)]$, and substitute for Π^I .

(ii) As $\delta \rightarrow 0$, (4.3) approaches $0 \leq \pi(N, D) - \pi(N/b, D)$, which is the defining condition for $D^*(N/b)$ in Proposition 3.1.

(iii) As T becomes large, the LHS of (4.3) approaches zero, while the RHS is unchanged. If D exceeds $D^*(N/b)$, then the first term on the RHS of (4.3) is nonnegative (by the definition of $D^*(N/b)$), so (4.3) holds for T sufficiently large. If D is less than $D^*(N/b)$, the RHS of (4.3) is positive if and only if $1 - \delta < \pi(N/b, D)/\pi(N, D)$. Using $\pi(n, D) = n \cdot m(n, D)$, this is the same as $\delta > 1 - b^{-1} \cdot m(N/b, D)/m(N, D)$, so if this holds and T is sufficiently large, (4.3) holds. Q.E.D.

REFERENCES

- 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. (1986). 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, Massachusetts.
- Besanko, D., and Thakor, A. (1993). Relationship banking, deposit insurance, and portfolio choice, in "Capital Markets and Financial Intermediation," (C. Mayer and X. Vives, Eds.), pp. 292–319. Cambridge Univ. Press, Cambridge, U.K.
- Bhattacharya, S., and Thakor, A. (1993). Contemporary banking theory, *J. Finan. Intermediation* 3, 2–50.

- Bordo, M., Redish, A., and Rockoff, H. (1995a). "A Comparison of the Stability and Efficiency of the Canadian and American Banking Systems, 1870–1925," NBER Historical Paper No. 67.
- Bordo, M., Redish, A., and Rockoff, H. (1995b). A comparison of the stability and efficiency of the Canadian and American banking systems in the Twentieth Century: Stability v. efficiency? in "Anglo–American Financial Systems: Institutions and Markets in the Twentieth Century" (M. Bordo and R. Sylla, Eds.), pp. 11–40. Irwin Professional Publishing, New York.
- Boyd, J., and Prescott, E. (1986). Financial intermediary coalitions, *J. Econ. Theory* **38**, 211–232.
- Calomiris, C. (1993). Regulation, industrial structure, and instability in U.S. banking: An historical perspective, in "Structural Change in Banking," (M. Klausner and L. White, Eds.), pp. 19–116. Business One Irwin, Homewood, IL.
- Campbell, T., and Kracaw, W. (1980). Information production, market signalling, and the theory of financial intermediation, *J. Finance* **35**, 863–882.
- Capie, F. (1995). Prudent and stable (but inefficient?): Commercial banks in Britain, 1890–1940, in "Anglo–American Financial Systems: Institutions and Markets in the Twentieth Century" (M. Bordo and R. Sylla, Eds.), pp. 41–64. Irwin Professional Publishing, New York.
- Capie, F., and Rodrik-Bali, G. (1982). Concentration in British banking, 1870–1920, *Business History* **24**, 280–292.
- Cerasi, V., and Daltung, S. (1995). "The Optimal Size of a Bank: Costs and Benefits of Diversification," working paper, Sveriges Riksbank.
- Chari, V., and Jagannathan, R. (1988). Banking panics, information, and rational expectations equilibrium, *J. Finance* **43**, 749–761.
- Checkland, S. (1975). "Scottish Banking: A History, 1695–1973," Collins, London.
- Cockerell, H., and Green, E. (1994). "The British Insurance Business: A Guide to its History and Records," Sheffield Academic Press, Sheffield.
- Collins, M. (1988). "Money and Banking in the UK: A History," Routledge, London.
- Diamond, D. (1984). Financial intermediation and delegated monitoring, *Rev. Econ. Stud.* **51**, 393–414.
- Diamond, D., and Dybvig, P. (1983). Bank runs, deposit insurance and liquidity, *J. Polit. Economy* **91**, 401–419.
- Dowd, K. (1992). Free banking in Australia, in "The Experience of Free Banking" (K. Dowd, Ed.), pp. 48–78. Routledge, London.
- Emmons, W. (1993). "Deposit Insurance and Last-Resort Lending as Delegated Monitoring: A Theory of Bank Safety Nets," working paper, Dartmouth College.
- Farrell, J., and Saloner, G. (1986). Installed base and compatibility: Innovation, product preannouncements, and predation, *Am. Econ. Rev.* **76**, 940–955.
- Flood, M., and Kwan, S. (1995). "The Boom and Bust Phenomenon and U.S. Bank Failure Experience, 1914–1929," working paper, Concordia University.
- Gehrig, T. (1995). "'Excessive Risk' and Banking Regulation," working paper, University of Basel.
- Gilson, S., John, K., and Lang, L. (1990). An empirical study of private reorganization of firms in default, *J. Finan. Econ.* **27**, 315–353.
- Gorton, G., and Pennacchi, G. (1990). Financial intermediaries and liquidity creation, *J. Finance* **45**, 49–71.

- Griffiths, B. (1973). The development of restrictive practices in the U.K. monetary system, *Manchester Sch. Econ. Soc. Stud.* **41**, 3–18.
- Innes, R. (1990). Limited liability and incentive contracting with ex-ante action choices, *J. Econ. Theory* **52**, 45–67.
- Jacklin, C., and Bhattacharya, S. (1988). Distinguishing panics and information-based bank runs: Welfare and policy implications, *J. Polit. Econ.* **96**, 568–592.
- Jenkins, D. (1984). The practice of insurance against fire, 1750–1840, and historical research, in “The Historian and the Business of Insurance” (O. Westall, Ed.), pp. 9–38. Manchester Univ. Press, Manchester.
- Katz, M., and Shapiro, C. (1985). Network externalities, competition, and compatibility, *Am. Econ. Rev.* **75**, 424–440.
- Krasa, S., and Villamil, A. (1992). A theory of optimal bank size, *Oxford Econ. Pap.* **44**, 725–749.
- Marcus, A. (1984). Deregulation and bank financial policy, *J. Banking Finance* **8**, 557–565.
- Matutes, C., and Vives, X. (1996). Competition for deposits, fragility, and insurance, *J. Finan. Intermediation* **5**, 184–216.
- Myers, S. (1977). Determinants of corporate borrowing, *J. Finan. Econ.* **5**, 147–175.
- Neufeld, E. (1972). “The Financial System of Canada,” St. Martin’s Press, New York.
- Rajan, R. (1992). Insiders and outsiders: The choice between informed and arm’s-length debt, *J. Finance* **47**, 1367–1400.
- Ramakrishnan, R., and Thakor, A. (1984). Information reliability and a theory of financial intermediation, *Rev. Econ. Stud.* **51**, 415–432.
- Roe, M. (1997). “Backlash,” Center for Law and Economic Studies Working Paper No. 127, Columbia University.
- Rothschild, M., and Stiglitz, J. (1970). Increasing risk: I. A definition, *J. Econ. Theory* **2**, 225–243.
- Shaffer, S. (1993). A test of competition in Canadian banking, *J. Money, Credit, Banking* **25**(1), 49–61.
- Sheppard, D. (1971). “The Growth and Role of UK Financial Institutions, 1880–1962,” Methuen, London.
- Smith, B. (1984). Private information, deposit interest rates, and the “stability” of the banking system, *J. Monet. Econ.* **14**, 293–317.
- Tirole, J. (1988). “The Theory of Industrial Organization,” MIT Press, Cambridge, MA.
- 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. (1995). Delegated monitoring and bank structure in a finite economy, *J. Finan. Intermediation* **4**, 158–187.
- Yanelle, M. (1989). The strategic analysis of intermediation, *Eur. Econ. Rev.* **33**, 294–301.
- Yanelle, M. (1997). Banking competition and market efficiency, *Rev. Econ. Stud.* **64**, 215–239.
- Yosha, O. (1997). Diversification and competition: Financial intermediation in a large Cournot–Walras economy, *J. Econ. Theory* **75**, 64–88.