On the Size Distribution of Banks

Huberto M. Ennis

In recent years, important changes to the U.S. banking regulatory framework have been introduced that were expected to affect the size distribution of banks. These changes in regulation had a clear objective: to allow for a higher degree of horizontal and vertical integration in the banking industry. While horizontal integration takes place when different firms that are producing the same product merge, vertical integration takes place when firms producing certain inputs merge with the firms that use those inputs.

The Riegle-Neal Interstate Banking and Branching Efficiency Act was passed in September 1994. The act allows banks and bank-holding companies to freely establish branches across state lines. In fact, the act came as the final step in a long process of gradual removal of interstate branching restrictions that took place at the state level during the late eighties and early nineties. This new flexibility in the branching regulation has opened the door to the possibility of substantial geographical consolidation in the banking industry. Indeed, geographical consolidation has always been one of the main channels used to achieve *horizontal* integration at an industry level.

In November 1999 Congress passed the Gramm-Leach-Bliley Financial Services Modernization Act. It allows affiliations among banks, securities firms, and insurance companies, removing many long-standing restrictions over the *horizontal and vertical* integration of firms providing financial services.

Both of these regulatory changes were expected to have substantial effects on the overall structure of the U.S. banking sector, and in particular, on the size

Research Department, Federal Reserve Bank of Richmond, huberto.ennis@rich.frb.org. The author wishes to thank Kevin J. Scotto for excellent research assistance and Ned Prescott for useful conversations on the subject. Kartik Athreya, Tom Humphrey, Pedro M. Oviedo, John Weinberg, Alex Wolman, Jose Wynne, and the seminar participants at NCSU and Duke University provided useful comments on a previous draft. All errors are my own. The views expressed here do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

distribution of banks. Some of these effects are already apparent in the data, and there may be more to come. It is not yet clear if the transition period is over. The question of whether *all* banks will eventually become nationwide banks is still very much unanswered. In other words, is there something special that community banks do which nationwide banks cannot replicate, or are small regional banks simply a consequence of long-lasting and strict government regulations? Even seven years after passage of the Riegle-Neal Act, there are still 7,920 small commercial banks (with less than a billon dollars in assets) representing 95 percent of the total number of banks in the system and holding 20 percent of total deposits. At the same time, there are 82 banks with more than $10 billon in assets that hold 70 percent of total deposits. These statistics indicate that even though some very large banks have already emerged, there are still many small banks with substantial participation in the administration of deposits.

In this article I will present some empirical and theoretical elements that could be used to support the view that the existence of community banks is justified even in an unregulated environment. Although the evidence is still preliminary, some interesting insights about the determinants of the banking industry structure arise from the discussion and can provide guidance for evaluating the future evolution of this important sector of the U.S. financial system.

The objective of the article is twofold. In Section 1, I will review stylized facts associated with the U.S. size distribution of banks and its evolution over the last 25 years. I will also include a brief discussion of the recent changes in U.S. regulation. Then, in Section 2, I will review some theoretical explanations for the coexistence of small and large banks in a competitive unregulated system. Section 3 provides conclusions.

1. SOME STYLIZED FACTS

Review of the Regulation

The Riegle-Neal Act is the final stage of a long process of bank branching deregulation in the United States. In 1975, no state allowed out-of-state bank holding companies to buy in-state banks, only 14 states permitted statewide branching, and 12 states completely prohibited branching. The rest had partial restrictions (for example, in some states a bank could only open branches in the county of its headquarters or in contiguous counties). These restrictions date from the Banking Act of 1933. However, starting in the late 1970s and continuing through the 1980s, all states relaxed their restrictions on both statewide and interstate branching (see Jayaratne and Strahan [1997, Table 1] for a list of the specific dates). Finally, in 1994 the Riegle-Neal Act removed all remaining restrictions on branching throughout the country.

It is probably safe to say that by the mid-1970s, the shape of the size distribution of banks fully reflected the effects of the branching restrictions that had been introduced 40 years earlier. Furthermore, the movement towards removing those restrictions in the 1980s surely explains the subsequent evolution of the distribution.

In the last decade, there has been a strong trend towards higher asset concentration in the industry.[1] One way to explain this trend is to acknowledge that the branching restrictions were probably highly binding while in place. Another and perhaps more interesting explanation is that the trend towards concentration appeared during a period when important technological innovations developed. There is little doubt that technological changes like computers and ATMs can help explain the observed increase in bank asset concentration. In fact, the potential efficiency gains associated with becoming a large high-tech bank may actually explain the political pressure for deregulation (see Broaddus [1998]). Deregulation is, to some degree, an endogenous event.

The fact that deregulation and technological innovation happened simultaneously has made it difficult to disentangle the independent effects of each of these factors on the size distribution of banks. Deregulation was a necessary condition for concentration, but probably not a sufficient one.

In 1999, the U.S. Congress passed another important piece of legislation that may strongly affect the market structure of the banking industry. The Gramm-Leach-Bliley Act created a new institution, the financial holding company, and allowed this new entity to offer banking, securities, and insurance products under one corporate roof. The law is still too recent to allow us to evaluate its long run impact on the financial services industry. However, two years after the law's enactment, there are a large number of banks that have taken advantage of the resulting opportunities for horizontal and vertical integration. Indeed, as of July 2001, 558 financial holding companies have been formed and 19 of the 20 largest banks in the United States now belong to a financial holding company.

The Gramm-Leach-Bliley Act also has provisions intended to increase competition and efficiency in the industry. Making an industry more competitive and efficient can change the flows of entry and exit, the optimal scale of operation, and the possibilities of growth at the firm level. These changes may in turn reshape the long-run size distribution of the surviving firms. However, it is still too early to conduct any conclusive evaluation of the actual effects of these provisions.

---

[1] As of March 2001, there were 18 commercial banks with more than $50 billon in assets; 8 of them had more than 1,000 branches in the United States. (The largest commercial bank, Bank of America, had more than $500 billon in assets and over 4,500 branches in the United States.)

There is another feature of the regulatory environment that can have important implications for the observed size distribution of banks. If the participants in the financial system have the perception that there exists a "too-big-to-fail" bias in the way regulators treat large institutions, then the level of asset concentration in the industry will tend to be higher and the size distribution more skewed to the right (with a disproportionately long right tail). Being a large institution presumably increases the ability of a bank to access the implicit subsidy associated with a too-big-to-fail policy. The existence of this type of policy in the U.S. banking industry is the subject of an ongoing debate (see, for example, Feldman and Rolnick [1997]). Furthermore, and most important for this article, it seems that isolating the effects of this particular policy over the scale of operation of banks can be a very complicated enterprise.

Data

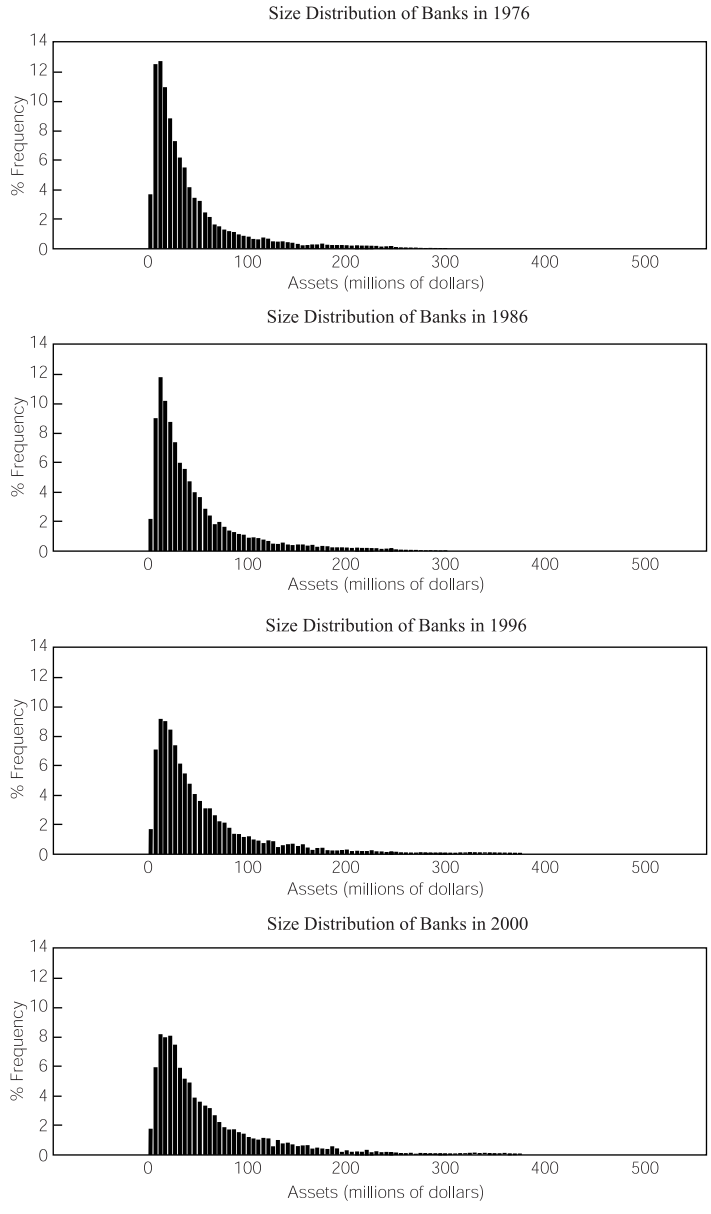
I now will present some statistics to characterize the size distribution of commercial banks in the United States and its evolution since 1976.[2] I use total assets to proxy the size of each bank, and all values are in real terms (dollars of 1982–1984). Figure 1 presents the histogram for the bottom (smallest) 95 percent of the total number of banks in each of four years 1976, 1986, 1996, and 2000. There is a wide range of bank sizes in each year. The distribution has shifted substantially in the last two decades. The average size has more than doubled (see Table 1). The density (frequency) of very small banks has clearly diminished.

Although there is a large number of small banks, the concentration in the industry is relatively high. Asset concentration has also increased in recent years. In Table 1, I report a time series for the Gini Coefficient of the asset size distribution in the industry.[3] The Gini Coefficient is relatively stable during the 1980s (with a value of around 0.84), but increases substantially after 1993 (reaching 0.90 in 2000). A noteworthy observation is that the density of midsize banks has increased. An important factor to have in mind when interpreting this fact is that the total number of banks in the system has been diminishing in the last decade, which means that higher densities do not imply a larger number of banks in certain ranges of the distribution. Figure 2 presents the histograms for banks with less than $400 million in assets (13,452 banks in 1986 and 7,745 in 2000). There seems to be a significant shift of the

[2] The data used here are from the Federal Reserve Bank of Chicago website (http://www.chicagofed.org/economicresearchanddata/data/bhcdatabase/subfiles.cfm).

[3] The Gini Coefficient is a measure of the degree of concentration associated with a given distribution of assets in the industry. It would be approximately equal to one when only 1 percent of the banks (the large banks) hold 99 percent of the assets in the industry and approximately equal to zero when all banks are of the same size.

Figure 1 Histogram of Bank Sizes (by Total Assets) (I)



mass of banks towards the right end of the distribution (although the absolute number of banks has been falling for almost all categories). In other words, compared with the size distribution of banks 20 years ago, today's distribution

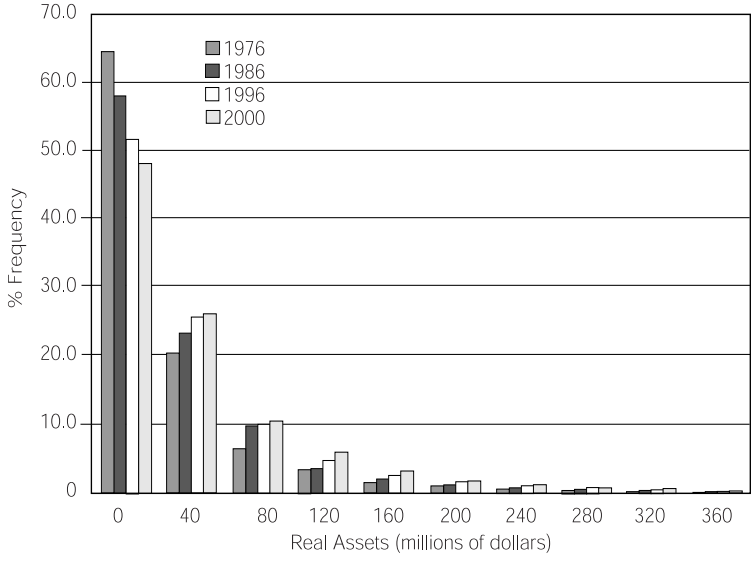
Table 1 Asset Concentration (I)

Year	Gini Coef.	Std. Dev. (Mean)	Number of Banks
1976	0.82	1,828 (140)	14,419
1977	0.83	1,965 (149)	14,417
1978	0.83	2,050 (154)	14,392
1979	0.83	2,077 (153)	14,356
1980	0.83	1,998 (149)	14,426
1981	0.83	1,953 (149)	14,407
1982	0.83	1,959 (155)	14,430
1983	0.83	1,877 (160)	14,420
1984	0.83	1,854 (165)	14,388
1985	0.84	1,921 (174)	14,278
1986	0.84	1,996 (188)	14,059
1987	0.84	1,932 (190)	13,553
1988	0.85	1,889 (199)	12,982
1989	0.85	1,933 (207)	12,572
1990	0.85	1,867 (206)	12,212
1991	0.85	1,883 (209)	11,807
1992	0.84	2,017 (216)	11,363
1993	0.85	2,188 (232)	10,881
1994	0.86	2,509 (256)	10,381
1995	0.87	2,749 (282)	9,875
1996	0.88	3,318 (302)	9,465
1997	0.89	4,055 (339)	9,081
1998	0.89	4,679 (377)	8,713
1999	0.90	5,476 (395)	8,520
2000	0.90	5,861 (427)	8,252

The mean and the standard deviation are in millions of 1982–1984 dollars.

has relatively fewer small banks, and, conditional on being small, banks tend to be larger today than in the past. It is not the case, then, that the very small banks disappearing in large numbers are losing all their market share to the extremely large national institutions. Intermediate-size banks are becoming relatively more important, too. In fact, the reduction in the number of small banks is especially concentrated on banks with less than 120 million dollars in assets, accounting for more than 96 percent of the reduction in the number of banks with less than 400 million (from 12,060 in 1986 to 6,558 in 2000). However, this shift in the relative mass of banks could be a consequence of the

Figure 2 Histogram of Bank Sizes (II)



transition process if very small banks are easier to take over and medium-size banks are simply in transition on their way to becoming larger institutions.

Table 2 further documents the level of concentration in the industry and its evolution over time. Again the table shows that concentration was stable (or slightly increasing) during the eighties and the early nineties and has significantly increased in the second half of the nineties. It is striking to note that the top 1 percent of the banks in the year 2000 own almost 70 percent of the assets (and the top 10 percent own almost 90 percent).

Table 3 presents some measures of the skewness (or asymmetry) of the distribution. In a symmetric distribution, the mean is located at the 50th percentile and the ratio of the mean to the median is 1. The bigger the concentration of assets in a few large banks, the more skewed to the right is the distribution. Indeed, according to the indicators in Table 3, the skewness of the asset distribution of banks has increased substantially during the nineties.

To try to determine the effect of government branching restrictions on the size distribution of banks, one can compare the distribution at the national level with that of a large state like California (Berger, Kashyap, and Scalise 1995). California has had no restrictions on statewide branching since the year 1909. The Gini Coefficient for the size distribution of banks in California was around 0.9 for most of the eighties and nineties, and the percentile location of the mean was around 94 percent. In summary, the concentration and the

Table 2 Asset Concentration (II)

Year	% of Assets (largest 1% of banks)	% of Assets (largest 10% of banks)	Ratio of largest 1% to smallest 40%	Ratio of largest 10% to smallest 40%
1976	55.8	78.1	15.6	21.8
1977	56.0	78.2	15.8	22.1
1978	56.8	78.7	16.4	22.7
1979	58.1	79.3	17.3	23.6
1980	58.1	79.4	17.1	23.4
1981	57.9	79.3	16.9	23.1
1982	57.3	79.2	16.8	23.2
1983	55.9	78.8	16.0	22.6
1984	55.6	79.0	16.2	23.1
1985	55.5	79.7	16.8	24.1
1986	55.4	80.1	17.2	24.8
1987	55.1	80.6	17.5	25.6
1988	54.7	81.1	18.0	26.8
1989	54.6	81.4	18.6	27.8
1990	54.1	81.3	18.2	27.3
1991	53.6	81.2	17.7	26.8
1992	54.0	81.1	17.6	26.5
1993	55.3	82.1	18.9	28.1
1994	56.7	83.5	21.2	31.2
1995	57.3	84.2	22.8	33.4
1996	60.9	85.0	25.8	36.0
1997	66.5	86.4	31.1	40.4
1998	68.0	87.2	33.8	43.4
1999	68.5	87.5	35.5	45.3
2000	70.2	88.2	38.6	48.5

skewness of the size distribution of banks in California during the eighties and nineties was very similar to that observed today for the national numbers.[4]

It is worth mentioning that using California as a benchmark for comparison became a less meaningful exercise after the mid-nineties deregulation of interstate branching. Indeed, in the last three or four years, changes at the national level have had some important indirect effects at the state level. Those

[4] During the seventies, bank-asset concentration in California was even higher (with a Gini Coefficient of around 0.94).

Table 3 Skewness

Year	Percentile Location of Mean	Ratio of Mean to Median
1976	90.6	4.9
1977	90.5	4.9
1978	90.8	5.0
1979	91.2	5.1
1980	91.3	5.2
1981	91.4	5.1
1982	91.4	5.1
1983	91.1	5.0
1984	90.8	5.1
1985	91.0	5.3
1986	91.1	5.4
1987	91.2	5.5
1988	91.3	5.8
1989	91.4	5.9
1990	91.2	5.9
1991	91.2	5.9
1992	91.3	5.8
1993	91.8	6.1
1994	92.1	6.6
1995	92.2	6.9
1996	92.7	7.3
1997	93.4	8.1
1998	93.8	8.5
1999	93.9	8.9
2000	94.2	9.5

effects were not present previously because the branching restrictions made California an isolated market.[5]

The histograms of bank sizes presented in Figure 1 resemble the probability distribution of a lognormal random variable. The lognormal distribution has been important in theoretical and empirical research. One of the most influential theories of the size distribution of firms was introduced by Robert

[5] In recent years, the measures of concentration and skewness for California have suffered large swings due to the fact that large state banks have merged with out-of-state banks and, in the process, have changed the location of their headquarters. (For example, from 1998 to 1999 the Gini Coefficient dropped from 0.92 to 0.84.)

Gibrat in the 1930s (see Sutton [1997]). His theory delivers a precise prediction for the long-run distribution of firm sizes: the lognormal distribution. Two strong assumptions are behind this prediction: (1) the number of firms is stationary and (2) the rate of growth of firms is given by an i.i.d. random variable independent of firm size. If one is willing to accept these assumptions as providing a reasonable representation of the evolution of a particular industry, then one can expect that the distribution of firm sizes will converge to the lognormal distribution.[6] Additionally, the lognormal distribution is a very convenient tool for analytical work. If a variable is lognormal, then the logarithm of that variable has a normal distribution. This means that a simple transformation of the data allows the researcher to apply all the well-known results associated with the normal distribution.

Because of the potential importance of lognormality, in Table 4 I perform some preliminary tests to see how far the U.S. commercial bank data is from delivering the lognormal distribution. The match is not very promising. The distribution of the logarithm of bank asset-size is relatively skewed to the right and has a higher degree of kurtosis (fatter tails or higher "peakedness," or both) than the normal distribution. Since the number of observations for each year is very large (around 10,000) we can safely conclude that these differences are not associated with sampling error: the distributions are significantly different. However, it should be said that during the years under consideration the industry has experienced important changes, and these calculations are not really appropriate as a test of Gibrat's theory (for that we would have to somehow control for the large flow of exit that took place in the industry).

On a related point, Simon and Bonini (1958) show that firm-entry assumptions matter for the determination of the stationary distribution. In particular, they combine Gibrat's firm-growth proportionality assumption with the assumption that new small firms enter the industry at a constant rate, and they show that the long-run size distribution approaches the Yule distribution (which has a fatter right tail than the log-normal).

2. SOME THEORETICAL EXPLANATIONS

There is an extensive literature on the size distribution of business firms that goes back to Gibrat's work during the 1930s. The literature on the size distribution of financial firms, however, is much smaller. In this section, I first

[6] This is actually not hard to see. Denote the size of the firm by x_t and let the i.i.d. random variable ε_t be a proportional rate of growth of the firm size. Then, we have that

$$\log x_t = \log x_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_t,$$

and the distribution of $\log x_t$ converges to the normal distribution as $t \to \infty$.

Table 4 Skewness and Kurtosis of the Log Data

Year	α_3 (Skewness)	α_4 (Kurtosis)	Ratio of Mean to Median
1976	1.02	5.80	1.0106
1986	1.13	6.07	1.0123
1996	1.23	6.52	1.0139
2000	1.25	6.89	1.0140
Normal Distribution	0.00	3.00	1.0000

The statistic $\alpha_3 = \mu_3/(\mu_2)^{3/2}$ and $\alpha_4 = \mu_4/(\mu_2)^2$ where μ_i is the ith moment about the mean, which is given by $\mu_i = (1/N)\sum_{j=1}^{N}(x_j - \mu)^i$ (μ is the mean of the distribution of x_js, N is the total number of banks and x_j is the asset size of bank j).

discuss in some detail one possible theory of bank size heterogeneity and then review some complementary theories available in the literature.

Explaining the size distribution of banks is a challenging task. There is always the possibility of extending the explanations used for business firms to the financial sector. Indeed, several of these theories are probably useful for explaining some of the size heterogeneity observed in the banking industry. But it seems that these theories will always be partial insofar as they do not recognize that there are some special characteristics of financial firms that act as the essential determinants of the size distribution of banks. One of these special characteristics is that banks play a role as information managers in the provision of credit. In the next subsection, I present a formal model that uses this characteristic to deliver a theory of the equilibrium heterogeneity of bank sizes.

The more traditional theories of firm size heterogeneity are founded on the notion of an underlying life cycle of firms.[7] The idea is that firms tend to be small at birth, after which they experience partially stochastic growth. This process generates a level of size heterogeneity in the long run that is not too far from the one observed in the business firm data. However, in this view, there is nothing special that small banks are doing which makes them different from large banks; they are just in the process of growing. This description does not seem to be a good representation of the U.S. banking industry, in which there is a large number of small banks that are not growing substantially through time and have no apparent intention of growing. The model presented next tries

[7] See Jovanovic (1982), Hopenhayn (1992), and Ericson and Pakes (1995). These models are modern versions of the traditional Gibrat's theory. They endogenize the growth process of firms and the decision of entry and exit.

to capture this point. It represents an economy where there are two different ways of organizing the production of information services by banks (one with small local banks and the other with large national banks) and these two organizational practices can coexist in equilibrium. In the second subsection, I discuss some alternative explanations of bank size heterogeneity that, in principle, should be taken as complementary to the formal model presented in the first part.

A Simple Model of the Size Distribution of Banks

I study an environment where two types of banks can coexist in equilibrium. On one side there is a large, geographically diversified national bank, with high leverage ratio (i.e., low bank-equity capital), and on the other side there are several small community banks restricting operation to one geographical area (hence not well diversified) and with lower leverage ratios.

The main motivation for the existence of banks in this model is their ability to monitor the behavior of investors with financial needs. Several other possible banking functions, including mobilizing funds and pooling risks, do not play a role in the present model. Banks can monitor investors, but monitoring is costly and not observable by third parties. If the bank is not well diversified, then it has to commit some of its own funds (i.e., hold some capital) so that depositors will become confident that the bank will perform the required monitoring activities. Because of this need for own funds, and because there are some fixed costs associated with becoming a bank, only wealthy individuals choose to become community bankers. The national bank, on the other hand, is well diversified and its owner does not need to commit his or her own funds to the operation. However, running a large institution involves some extra operational costs. Because of the economies of scale associated with the fixed cost of setting up a bank, only one diversified bank exists in equilibrium. Having only one national bank is an extreme situation but of no fundamental importance for the points that I intend to illustrate with the model. A minor extension of the model would allow for the existence of several large banks in equilibrium (for example, by introducing managerial ability, as in Lucas [1978]).[8]

The main idea underlying the model is that there are two possible ways to provide a specific service (in this case, management of information). One way is to run a community bank with high capital ratios and low operating costs and the other way is to run a national bank with low capital ratios and

[8] Another way to generate a bounded optimal size of the diversified banks is to assume that the average cost of monitoring, constant in the present article, is instead increasing in the size of the bank (see Cerasi and Daltung [2000]).

higher operating costs. Both ways can be made equally efficient and hence can coexist in equilibrium.

Two interesting results emerge from the comparison of the equilibria when there is a national bank in the system and when national banks are exogenously ruled out (for example by regulation). First, lower levels of total investment are observed in the equilibrium with no national bank. Second, in the equilibrium with a national bank there are fewer community banks and they tend to be smaller in size. Some of these facts are consistent with the evolution of the U.S. banking industry after branching deregulation (see Section 2).

I turn now to the details of the model.[9] Assume that there are a large number of different geographic (or economic) zones in the economy. There is a continuum of risk-neutral investors living in each zone. For simplicity I assume the population of investors in each zone has size 1. Investors are indexed by the amount of funds they own. Let $\widetilde{\theta} \in [0, 1]$ be the amount of funds owned by investor $\widetilde{\theta}$. We also assume that there is only one investor for each level of $\widetilde{\theta}$. A more general assumption would be needed to obtain a realistic size distribution of banks. At the beginning of the period, agents have to invest (or store) their funds in order to have them back at the end of the period when they will be used to pay for consumption.

Each zone has available a large number of risky investment projects. Each project is associated with an entrepreneur and, when undertaken, can either succeed or fail. We index projects by their productivity when success occurs, $r_A \in [1, 2]$, and projects are uniformly distributed across the possible values of r_A. Success and failure are verifiable, but the value of r_A is private information to the entrepreneur. When the project fails, the return is zero. In other words, project r_A has productivity r_A when success happens. Each project is owned by an entrepreneur that can choose to exert effort in running the project. A project requires I units of funds to be undertaken. If the project is undertaken with effort, the probability of success is given by p_H. The probability of success for projects undertaken with no effort is p_L. We assume that p_H is greater than p_L. Projects within a zone are perfectly correlated (they all fail together) and projects in different zones are independent.[10] We assume that for a project to be undertaken with effort, it has to be monitored by a bank. Finally, assume that only projects undertaken with effort can have a positive net present value. Hence, the incentive compatible allocation is the unique implementable allocation. Assume, for simplicity, that there is a given gross

[9] The model shares some similarities with those used in Holmstrom and Tirole (1997) and Ennis (2001).

[10] In equilibrium all projects will be undertaken exerting effort. The underlying assumption on success correlation is that projects undertaken with no effort fail when projects undertaken with effort fail, as well as some other times (so that $1 - p_L > 1 - p_H$). See Holmstrom and Tirole (1997, footnote 8) for details.

Table 5 Notation

$\widetilde{\theta}$	funds owned by investor $\widetilde{\theta}$
R	gross safe interest rate
r_A	return of project r_A when success
p_H (p_L)	prob. of success with (without) effort
I	size of investment projects (in amount of funds needed)
c	cost of monitoring a project
κ	cost of becoming a bank
δ	cost of diversification
ψ	size of the community banks (number of projects monitored)
ψ_D	size of the diversified bank
I_m	bank capital per project
r_P	interest rate on deposits (deposit interest rate)
$\widehat{r}_A$ (r_A^*)	interest rate on bank loans with (without) branching restrictions
$\widehat{\theta}$ (θ^*)	funds owned by the smallest bank with (without) branching rest.
Θ	total amount of monitors' own funds

interest rate R on funds. We can think of R as the return obtained from a safe storage technology.[11]

Assume monitoring is costly and not observable. Let c be the per-project cost of monitoring. The cost c is in utility terms (it does not deplete available funds). Any investor in the economy can choose to become a monitor. For reasons that will become clear below we can call each of these monitors a bank. To acquire the monitoring technology the agent has to incur a cost κ (in utility terms). Given that an agent has incurred the cost κ, the agent can monitor as many projects as desired as long as he or she incurs the cost c per project being monitored. This makes the market for monitoring services perfectly competitive. The monitor can also choose whether to handle projects in one zone or in a large number of zones.[12] Assume that there is an extra operational cost δ of running an institution (bank) handling projects in more than one zone. Then, we need to consider only two possible levels of diversification: the monitor either specializes in projects from one particular zone or becomes completely diversified.

[11] The following restrictions on fundamental parameters are assumed to hold: $2p_L < R < R + c/I < 2p_H$ and $p_H < R + c/I$.

[12] Specifically we assume that there is a continuum of different zones with total measure of one. See Ennis (2001) for details.

Bank Branching Restrictions

Let us consider first the case where each monitor is exogenously restricted to handle projects from a single zone. An agent with a monitoring technology accepts funds from other agents and invests in projects. These external funds available to the monitor can be called deposits. If a bank only handles projects in one zone, then the bank fails with probability $1 - p_H$, the probability that the projects in the zone fail. Let r_p be the return on deposits when the bank does not fail. By an arbitrage condition we have

$$p_H r_p = R.$$

This condition means that the expected rate of return on deposits in a community bank is equal to the safe interest rate.

It is not hard to see that in equilibrium there is a threshold on the productivity of projects, $\widehat{r}_A$, such that only projects with $r_A \geq \widehat{r}_A$ will be undertaken. Consequently, $2 - \widehat{r}_A$ is the total number of projects undertaken. Because there is perfect competition in the market for monitoring services, the project owners only pay $\widehat{r}_A I$ to the bank in return for a loan of size I. For this reason we can call $\widehat{r}_A$ the loan interest rate. Let ψ be the number of projects handled by a bank. The variable ψ is an indicator of the size of the bank. Agents agree to deposit funds in a bank of size ψ only when the following incentive compatibility condition is satisfied

$$-c\psi + p_H \left[\widehat{r}_A I\psi - r_p(I - I_m)\psi\right] \geq p_L \left[\widehat{r}_A I\psi - r_p(I - I_m)\psi\right], \quad (1)$$

where I_m is the amount of own funds the bank commits per project. This condition says that the return to the banker from monitoring the projects must be greater than the return from not monitoring (given that depositors believe that the bank *will* be monitoring). Because monitors want to handle as many projects as possible, condition (1) holds with equality in equilibrium and determines the equilibrium bank capital per project, $\widehat{I}_m$. For this reason, the banker's return per project must satisfy

$$-c + p_H \left[\widehat{r}_A I - r_p(I - \widehat{I}_m)\right] = \frac{p_L c}{p_H - p_L}.$$

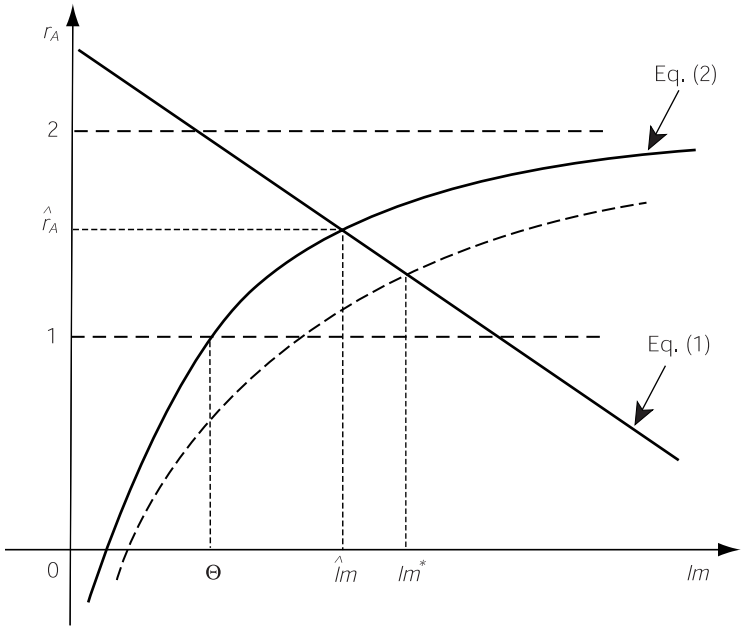
Let Θ be the total amount of own funds committed by monitors in equilibrium. The next paragraph explains how this quantity is determined. Remember that $(2-\widehat{r}_A)$ is the number (measure) of projects undertaken in equilibrium. Then, market clearing for the funds owned by monitors requires that

$$(2 - \widehat{r}_A)\widehat{I}_m = \Theta. \quad (2)$$

This states that the number of projects funded times the amount of the banker's own funds invested per project equals the total amount of banker's funds invested. Given Θ, we can use expressions (1) and (2) to determine the equilibrium values of $\widehat{r}_A$ and $\widehat{I}_m$ (see Figure 3).

Consider now the decision of an investor to become a bank. Note that because of the incentive-compatibility constraint (1) for monitors, the return

Figure 3 Equilibrium Loan Interest Rate



The intersection of the incentive-compatibility constraint for community banks (equation (1)) and the market clearing condition for funds owned by monitors (equation (2)) determine the equilibrium loan interest rate. The dashed locus corresponds to the shift in equation (2) when a diversified monitor is introduced in the model. See Table 5 for notation.

associated with becoming a bank is given by

$$-\kappa + \left(-c + p_H\left[\widehat{r}_A I - r_p(I - \widehat{I}_m)\right]\right)\psi = -\kappa + \frac{p_L c}{p_H - p_L}\psi.$$

As long as the return from becoming a bank is greater than $R\widehat{I}_m\psi$ (the return from safely storing funds), the agent will choose to become a bank. Because the return is increasing with the number of projects monitored, there is a minimum equilibrium size of banks, $\widehat{\psi}$, determined by

$$-\kappa + \frac{p_L c}{p_H - p_L}\widehat{\psi} = R\widehat{I}_m\widehat{\psi}. \tag{3}$$

Since the amount of funds banks commit to each project, $\widehat{I}_m$, is uniform across projects, a particular value of ψ (the size of the bank) is directly associated with a particular value of the wealth of the banker, $\widetilde{\theta}$. This relationship is

given by the following equation

$$\psi = \frac{\widetilde{\theta}}{\widehat{I}_m}. \tag{4}$$

Then, given the value of $\widehat{\psi}$ that solves equation (3), there is a threshold on the amount of funds that an agent has to own in order to become a bank. Call this threshold $\widehat{\theta}$. All agents with $\widetilde{\theta} > \widehat{\theta}$ will become banks, and $1 - \widehat{\theta}$ is the total number of banks in each zone. The total amount of own funds invested by banks in equilibrium is then given by

$$\Theta = \int_{\widehat{\theta}}^{1} \widetilde{\theta} d\widetilde{\theta} = \frac{1}{2}\left(1 - \widehat{\theta}^2\right). \tag{5}$$

Substituting expressions (1) and (5) into equation (2) we obtain what can be thought of as a demand for bank funds, $\widehat{\theta}^d = f^d(I_m)$.[13] Equation (3) implicitly defines a supply of bank funds. By making demand equal supply, we can obtain the equilibrium level of $\widehat{\theta}$ (see Figure 4).

Note that this equilibrium induces a size distribution of banks (monitors) according to equation (4). These banks are all essentially the same type of institution (community banks). The size distribution is a direct consequence of an underlying heterogeneity among bank owners (in terms of own funds) that is exogenous to the model. In what follows we introduce some *endogenous* heterogeneity.

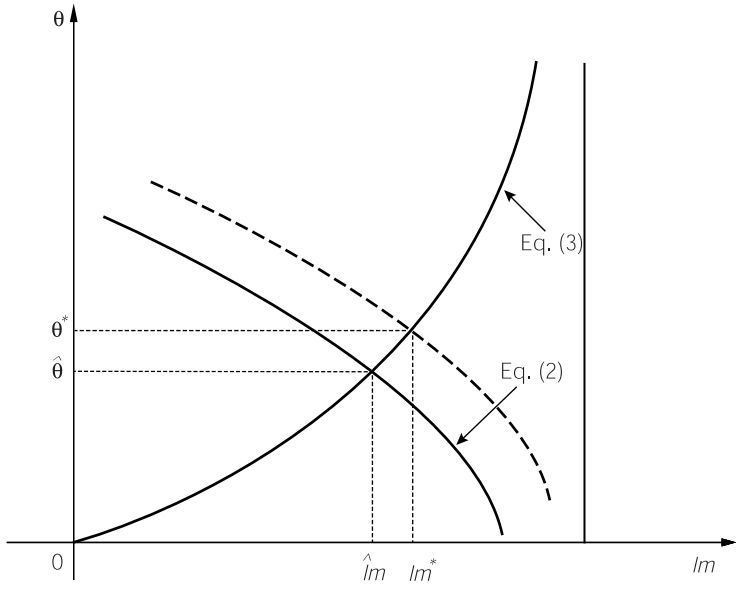
No Bank Branching Restrictions

Consider the case when full diversification is possible, i.e., when we do not restrict banks to handling projects in only one zone. Fully diversified monitors do not face an information problem. The proportion of failed projects ("bad loans") in a monitor's portfolio is observable by third parties, and this proportion reveals the bank's monitoring activities.[14] Anyone can become a well-diversified monitor; no internal funds are needed. However, because there is a fixed cost $\delta + \kappa$ of establishing a diversified bank, economies of scale imply that only one diversified bank will exist in equilibrium. We assume contestability and hence a zero profit condition must hold (see Tirole [1988], 307). Let ψ_D be the number (measure) of projects handled by the

[13] Note that the right-hand side of equation (2) gives us the amount of monitor funds needed to run $(2 - \widehat{r}_A)$ investment projects when $\widehat{I}_m$ monitors funds are needed per project to satisfy the incentive compatibility constraints.

[14] Note that diversification is not originated on risk aversion considerations. All agents are risk neutral in the model. See Diamond (1984) for a related result. In Diamond's model, diversification allows the economy to save on actual monitoring costs. In the model in this article, diversification allows the possibility of running a bank without committing internal funds. No saving of monitoring cost goes on here.

Figure 4 Demand and Supply of Bank Funds



The demand (equation (2)) and supply (equation (3)) of funds owned by monitors determine the number of community banks in the economy, $1-\widehat{\theta}$. The dashed locus represents the shift in equation (2) when a diversified monitor is introduced in the model.

diversified monitor. The zero profit condition is

$$\left[p_H\left(r_A - r_p\right) I - c\right]\psi_D - (\delta + \kappa) = 0. \tag{6}$$

This condition states that the net return per loan in the diversified bank multiplied by the size of the bank (i.e., the total number of loans in the bank) has to equal the fixed cost of setting up the diversified institution. The market clearing condition for monitors' funds is now given by

$$(2 - r_A - \psi_D) I_m = \Theta. \tag{7}$$

That is, the number of projects monitored by community banks multiplied by the amount of bank capital per loan has to equal the total amount of bankers' funds invested. Equations (1), (3), (4), and (5) still hold in equilibrium. For intermediate values of δ a well-diversified bank (monitor) will coexist with the community banks in equilibrium. For notational convenience, I use an asterisk to indicate the value of the variables in the equilibrium with a diversified bank and a hat for the equilibrium with no diversification.

The first important result is that the well-diversified bank is also a large bank, i.e., $\psi_D^* > \psi^*$ (where ψ^* is the size of the smallest community bank).

Table 6 Bank Size and Capital Ratios

Asset Size	1976	1986	1996	2000
$\leq$ 40 million	9.2	9.6	11.7	13.2
$\leq$ 50 billion	7.5	7.6	9.7	9.7
$>$ 50 billion	5.2	5.3	7.9	8.6

To see this, note that if we plug (3) and (6) into equation (1) (holding with equality) we obtain that

$$\frac{\kappa}{\psi^*} = \frac{\delta + \kappa}{\psi_D^*},$$

which implies that $\psi_D^*/\psi^* > 1$. Diversified banks must be larger in order to generate sufficient returns to cover the fixed cost, δ, of lending across different regions.

We turn now to comparing the value of some fundamental variables under the two possible cases: when diversification is ruled out exogenously and when it is not. Think of this comparison as a way to improve our understanding of the long-run implications associated with removing geographic (and possibly other) restrictions on the level of integration in the banking industry.

The second important result is that there are fewer single zone banks when a well-diversified bank is part of the system, i.e., $1 - \theta^* < 1 - \widehat{\theta}$. To see this, note that having $\psi_D > 0$ in equation (7) shifts the demand curve for community bank funds to the right (see Figure 4), increasing the equilibrium threshold to θ^*. It is worth noticing that the banks that are disappearing are the smallest (those for which $\widetilde{\theta} \in [\widehat{\theta}, \theta^*]$). In Figure 4 we can also see that $\widehat{I}_m < I_m^*$. This inequality is the foundation for the following two results.

The third result is that the number of projects undertaken in equilibrium is smaller when there is no diversified bank, i.e., $2 - \widehat{r}_A < 2 - r_A^*$. This result is a direct implication of the fact that equation (1) holds (with equality) in both equilibria and that $\widehat{I}_m < I_m^*$.

Finally, the fourth result is that non-diversified banks tend to become smaller in the equilibrium with a diversified institution. From equation (4), given a value of $\widetilde{\theta}$, a non-diversified bank will hold a smaller number of projects in the equilibrium with the larger I_m, that is, in the equilibrium with the diversified institution.

In terms of the implications for the observed size distribution of banks, using equation (3) we can see that $\psi^* > \widehat{\psi}$, i.e., the smallest community bank is larger when there is a diversified bank. When a diversified bank enters the market, the equilibrium loan interest rate (r_A^*) falls, reducing the profitability of community banks. As a consequence, only larger community banks survive.

Four final remarks seem relevant at this point. First, note that by adjusting the distribution of agents over the level of own funds $\widetilde{\theta}$, one can easily match any given size distribution of community banks. The assumption of a uniform distribution over $\widetilde{\theta}$ is convenient but is in no way necessary. Second, I have assumed that the market for national banks is contestable. This assumption allowed us not to worry about the possibility of monopoly power even though only one national bank exists in equilibrium. Contestability has been challenged on several grounds in the theoretical literature (see Tirole [1988], Chapter 8). The implications of increasing bank-asset concentration on the level of competition in the industry are of major concern to researchers and policymakers. I abstracted from these considerations in the model, but they are probably important and merit further study. Third, note that the model has implications for the amount of bank capital that community and national (diversified) banks would hold in equilibrium. Preliminary analysis of the data shows that, in accordance with the model, small community banks tend to systematically hold higher capital ratios than large national banks (see Table 6). Finally, the size of business firms plays no role in the model presented here. All investment projects are the same size and have the same financing requirements. Empirical studies tend to find that small firms rely more heavily on banks for their financing needs (compared with large firms). The model presented here is too simple to be used to study this last issue. However, below I discuss some complementary theories for which the size of business firms is important.

Other Theoretical Explanations of the Bank Size Distribution

Product Differentiation

It has been well documented that small businesses tend to rely heavily on bank credit (see Bitler, Robb, and Wolken [2001]). Small banks that maintain a long-term relationship with borrowers provide an important share of this credit. For example, Strahan and Weston (1996) document that the market share of small banks in the market for loans to small firms was 35 percent in 1995. This stylized fact can be used as a foundation for a product-differentiation theory of the size distribution of banks.

Banks provide differentiated financial services. For example, a bank could make available standardized loans, for which the approval procedure and the necessary monitoring are systematic and uniform across borrowers, or it could provide customized loan contracts to long-term clients. But, in principle, a single bank could also provide both. Some other factor needs to be introduced to explain the different sizes of banks. One possibility is that there are some technological reasons that make the provision of both types of loans by uniform

size banks inefficient. There are two issues related to this argument that need explaining. First, why are large banks more efficient at providing standardized loans and, second, why are small banks more efficient at relationship lending? The answer to the first question could be in the existence of economies of scope. Usually, standardized loans are more appropriate for large firms because the information required for the loan is more readily available and verifiable. At the same time, large firms tend to demand a wider array of products and services from the bank. In most cases, only large banks can satisfy all those demands efficiently (perhaps as a matter of being able to achieve the optimal scale of production).

The harder question is why large institutions cannot replicate the relationship lending practices of small banks. In fact, Strahan and Weston (1996) find that in 1995 large U.S. banks had a significant participation rate in the market for loans to small businesses (35 percent).[15] Perhaps the question should be rephrased in terms of the difference in bank portfolio shares of small business loans. In 1995, small commercial and industrial loans represented only 3 percent of total assets of large banks as opposed to 9 percent of small banks (see Strahan and Weston [1996]).

One possible explanation for this difference can be found in the combination of two factors: it is harder to monitor lending decisions in large banking organizations, and relationship loans require more discretion by loan officers. As a consequence of these two factors, small banks tend to be more efficient in the provision of this kind of loan. Regardless of the details, what supports this theory is the underlying heterogeneity of business firms. Because there is a size distribution of business firms, there is a size distribution of banks.

This theory, based as it is on a demand for differentiated products, also has implications for the interpretation of the recent changes in the U.S. banking industry. In the long run, a larger share of the market for loans to small firms will probably be held by large banks, but it is also likely that some small banks will continue to exist (due to their relative efficiency in the provision of relationship loans). Finally, it is important to highlight that, according to this theory, the evolution of the size distribution of business firms should directly affect the size distribution of banks. In other words, if technological developments drive the optimal scale of most business firms to become ever larger (Lucas 1978), then the role of small banks in the economy will also tend to decrease with time.

---

[15] In recent years the approach of large banks to small-business lending has experienced important changes. More and more large banks have started to adopt automated underwriting systems based on credit scoring. This allows large banks to make small business loans on a large scale. See Frame, Srinivasan, and Woosley (2001) for an updated account of this new development.

Corporate Governance

Issues in corporate governance of financial institutions are potentially important for explaining the size distribution of banks. Some authors have argued that internal corporate governance tends to be weaker for banks than for other types of corporations (see Prowse [1997]).[16] Here I sketch one theory of bank size that is based on these considerations. The idea is not to provide a definitive explanation of size heterogeneity but to illustrate how weak corporate governance may affect bank-size dispersion.

There are numerous empirical studies documenting that recent bank mergers do not seem to result in large efficiency gains (see, for example, Berger, Demsetz, and Strahan [1999]). The traditional justifications for mergers (for example, economies of scale and scope) have problems accounting for these findings. Some efforts have been made to provide alternative explanations for the tendency of banks to become large. One of these possible explanations is based on imperfect corporate governance and uncertainty about the managerial ability of bank CEOs (see Milbourn, Boot, and Thakor [1999]). This explanation can also provide justification for some of the bank size dispersion observed in the data. In fact, talent heterogeneity among bank CEOs alone could be used to induce a size distribution of banks, as in Lucas (1978). However, the corporate governance story involves information issues that were not present in Lucas's paper.

The main objective of the theory is to explain mergers that do not imply efficiency improvements, which is not so important to the present article; however, the theory's prediction of some size heterogeneity among banks is more germane. Suppose that shareholders do not know how talented the CEO of their bank is, but they would like to better compensate a talented CEO. Since talent is associated with a higher probability of success, the shareholders will use the success rate of the CEO as a proxy for his or her talent. However, not only talented CEOs are successful; some CEOs are just lucky. This brings up a problem: Inferring who is talented is not an easy task. Suppose further that as the bank gets bigger, it becomes harder for the CEO to just get lucky. CEOs who perceive themselves as talented individuals will then tend to prefer to manage large institutions (or make their institutions bigger by completing mergers and acquisitions) because if they eventually become successful, they will more clearly signal their ability and thereby increase their compensation. It can be shown that in this kind of environment, CEOs will tend to generate and manage different sizes of banks according to their perception of their own ability (not known to them with certainty).

[16] For example, government measures regulating bank takeovers, such as the need for prior approval and other potential delays, make the possibility of takeovers a less effective mechanism for disciplining bank managers.

An interesting extension of this theory suggests that there may be a bias towards large organizations in the banking industry. Suppose that the more talented CEOs tend in fact to perceive themselves as more talented (and hence to manage large banks). Suppose also that shareholders have this information and intend to use it in their compensation decisions. Less talented individuals may then choose to manage large institutions just to avoid revealing that they are actually not in the group of talented managers.[17]

3. CONCLUSIONS

This article provides an overview of some empirical and theoretical issues associated with the existence of a nondegenerate size distribution of banks in the United States. I review a number of theories of bank size heterogeneity, and I concentrate on those theories that tend to explain the small-banks phenomenon not as a transitory situation but as the result of an explicit equilibrium choice. This explanation seems to be in accord with the empirical facts described in the first part of the article. The size distribution of banks tends to be relatively more skewed to the right than life-cycle-of-firms theories predict. In other words, the mass of banks is highly concentrated around the range of small asset size. The theories reviewed in this article could help explain this fact.

But it is also true that 50 years of heavy regulation in the banking industry, and branching restrictions in particular, have played a major role in shaping the size distribution of banks in the United States. Deregulation is still very recent, and it may well be that the transition to a new banking industry structure is not over yet. For example, the banking system in Canada, which has never had branching restrictions, has mainly large banks with numerous branches across the country. The question remains, will the U.S. system converge to the Canadian model of banking? One possibility is that the final industry structure will be influenced by initial conditions even after the transition period is over. For example, community banks, having existed for some time, may have generated a demand for their services that will persist. If this is the case, then the market structure of the U.S. banking system and the Canadian system will continue to be different even after their regulatory frameworks have fully converged.

---

[17] Bliss and Rosen (2001) study the relationship between bank mergers and CEOs' compensation in a sample of megamergers that took place between 1986 and 1995. They find significant evidence supporting the hypothesis that asset growth (especially via mergers) tends to increase CEOs' compensation. They also find that this effect tends to motivate acquisition decisions by CEOs. (CEOs with a higher proportion of stock-based compensation tend to be less likely to engage in an acquisition.)

REFERENCES

Berger, Allen N., Anil K. Kashyap, and Joseph M. Scalise. 1995. "The Transformation of the U.S. Banking Industry: What a Long, Strange Trip It's Been." *Brookings Papers on Economic Activity* 2: 55–201.

_____, Rebecca S. Demsetz, and Philip E. Strahan. 1999. "The Consolidation of the Financial Services Industry: Causes, Consequences, and Implications for the Future." *Journal of Banking and Finance*. 23 (February):135–94.

Bitler, Marianne P., Alicia M. Robb, and John D. Wolken. 2001. "Financial Services Used by Small Businesses: Evidence from the 1998 Survey of Small Business Finances." *Federal Reserve Bulletin*. (April): 183–205.

Bliss, Richard T., and Richard J. Rosen. 2001. "CEO Compensation and Bank Mergers." *Journal of Financial Economics*. 61 (July): 107–38.

Broaddus, J. Alfred, Jr. 1998. "The Bank Merger Wave: Causes and Consequences." Federal Reserve Bank of Richmond *Economic Quarterly* 84 (Summer): 1–11.

Cerasi, Vittoria, and Sonja Daltung. 2000. "The Optimal Size of a Bank: Costs and Benefits of Diversification." *European Economic Review* 44 (October): 1701–26.

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

Ennis, Huberto M. 2001. "Loanable Funds, Monitoring and Banking." *European Finance Review* 5: 79–114.

Ericson, Richard, and Ariel Pakes. 1995. "Markov-Perfect Industry Dynamics: A Framework for Empirical Work," *Review of Economic Studies* 62 (January): 53–82.

Feldman, Ron J., and Arthur J. Rolnick. 1997. "Fixing FDICIA. A Plan to Address the Too-Big-To-Fail Problem." Federal Reserve Bank of Minneapolis *Annual Report*.

Frame, W. Scott, Aruna Srinivasan, and Lynn Woosley. 2001. "The Effect of Credit Scoring on Small-Business Lending." *Journal of Money, Credit and Banking* 33 (August): 813–25.

Holmstrom, Bengt, and Jean Tirole. 1997. "Financial Intermediation, Loanable Funds, and The Real Sector." *Quarterly Journal of Economics* 62 (August): 663–91.

Hopenhayn, Hugo A. 1992. "Entry, Exit, and Firm Dynamics in Long Run Equilibrium." *Econometrica* 60 (September): 1127–50.

Jayaratne, Jith, and Philip Strahan. 1997. "The Benefits of Branching Deregulation" Federal Reserve Bank of New York *Economic Policy Review* (December): 13–29.

Jovanovic, Boyan. 1982. "Selection and the Evolution of Industry," *Econometrica* 50 (May): 649–70.

Lucas, Robert E., Jr. 1978. "On the Size Distribution of Business Firms." *The Bell Journal of Economics* 9 (Autumn): 508–23.

Milbourn, Todd T., Arnoud W. A. Boot, and Anjan V. Thakor. 1999. "Megamergers and Expanded Scope: Theories of Bank Size and Activity Diversity" *Journal of Banking and Finance* 23 (February): 195–214.

Prowse, Stephen. 1997. "Corporate Control in Commercial Banks." *The Journal of Financial Research* 20 (Winter): 509–27.

Simon, Herbert A., and Charles P. Bonini. 1958. "The Size Distribution of Business Firms." *The American Economic Review* 48 (September): 607–17.

Strahan, Philip E., and James Weston. 1996. "Small Business Lending and Bank Consolidation: Is There Cause for Concern?" Federal Reserve Bank of New York *Current Issues in Economics and Finance* 2 (March): 1–6.

Sutton, John. 1997. "Gibrat's Legacy." *Journal of Economic Literature* 35 (March): 40–59.

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