Competition and Increasing Returns to Scale: A Model of Bank Size*

Tianxi Wang

Abstract

This paper examines the causal effects of bank size on banks' survival, asset quality, and leverage. Two forces drive these effects: increasing returns to scale derived from banks' expertise; and competition. The first enables bigger banks to survive competition better, have higher asset-quality, and be more leveraged. It drives banks into a race for expansion. This race toughens competition between banks, which edges out small banks and may deteriorate all banks' asset quality. Consequently, the banking industry will be dominated by a small number of highly leveraged banks. In this paper, financial intermediation arises endogenously and coexists with direct finance.

Key Words: Bank Size; Competition; Increasing Returns to Scale; Asset Quality; Leverage; Financial Intermediation

*Email: wangt@essex.ac.uk. Correspondence: Department of Economics, University of Essex, Colchester, CO4 3SQ, UK. Fax: +44 (0) 1206 872724. I am indebted to Dmitri Vinogradov and Giovanni Ko for their enormous help with the exposition, and to the anonymous referee and the editor, whose comments greatly improved the paper. I thank, for their helpful coments, Roy Bailey, Sanjay Banerji, Hans Gersbach, Xuewen Liu, John Moore, Jean Rochet, Zhen Song, Tuomas Takalo, Huainan Zhao, and seminar participants at Zhejiang University, Fudan University, Essex University, Bank of Finland, ETH Zurich, and Nottingham Business School.

Why is the banking industry dominated by a few "too big to fail" institutions? As a step towards answering the question, this paper considers the economic implications of bank size, which, unlike bank capitalization, receives little attention in the academic literature. Specifically, the paper examines the causal effects of bank size on banks' survival in a competitive market and on their asset quality. Based on these effects, the paper also considers how size affects banks' choice of leverage.

This paper characterizes the banking industry with two assumptions. First, banks have an advantage over the general public (households) in identifying which investment projects of entrepreneurs are profitable. In this paper, while households have no way to evaluate projects, banks can attain expertise to evaluate and screen them. Therefore, households invest in a project only if some bank certifies to them that it finds this project profitable. Second, banks can provide this certification not by words of mouth, but by investing a sufficient quantity of their own funds in the projects. Thus, informed funds – namely, those provided by banks – earn a higher rate of return than uninformed funds – namely, those of households.

These two assumptions imply that the banking industry has increasing returns to scale, which is consistent with the industry's long-term trend of increasing concentration.[1] If a bank has attained a certain level of screening expertise, it can apply this expertise to the deployment of all its funds. Thus, all its funds earn the return of informed funds. The more funds a bank has, the bigger profit it earns from its screening expertise, which induces the bank to attain a higher level of the expertise. Increasing returns to scale may, in general, result from any sort of expertise because the acquisition of expertise is in the nature of fixed costs. For example, learning a widely spoken language delivers

[1] This trend is well documented in the empirical literature; see, amongst others, Berger, Kashyap and Scalise (1995) (for US over 1979-1994), Saunders and Wilson (1999) (figures 2 and 3, for Canada and UK over 1893-1991), Berger, Demsetz and Strahan (1999) (for US over 1988-1997), Jones and Critchfield (2005) (for US over 1984-2003).

bigger benefits than learning a narrowly spoken one; and becoming a top comedian is more profitable where the potential audience is larger.[2] To exploit the increasing returns to scale, banks all want to expand. Then, competition becomes fiercer. The larger the quantity of informed funds banks supply, the lower the return rates these funds earn. Consequently, banks all gain less from screening expertise and thus attain less of it.

The two forces, increasing returns to scale and competition, are in conflict. Together they shape the implications of banks' sizes for their survival, asset quality, and leverage, and drive the banking industry to be dominated by a small number of highly leveraged banks.

First, a bank can survive competition only if its size is above a threshold, and this threshold is higher when bank finance is more abundant. This result is consistent with the previously mentioned trend of increasing concentration: the pressure to survive drives banks all to expand, which raises the survival threshold, edging smaller banks out.

Second, if all banks are enlarged, but their shares of the loan market remain fairly stable, then their screening expertise declines and their asset quality falls. This is because, for a bank whose market share is not much increased, the negative effects of competition dominate the positive effects of increasing returns to scale. What banks sell to entrepreneurs, essentially, is certification service rather than financing, because households have abundant funds. If a bank cannot sell the service to more entrepreneurs, the scale of its business is not enlarged and the force of increasing returns to scale not unleashed.

Third, there is a complementarity between leverage and screening expertise. On the one hand, the higher a bank's level of screening expertise, the greater its leverage. On the other hand, the greater leverage enlarges the scale of its funds more, which, due to increasing returns to scale, induces the bank to attain a higher level of screening expertise.

[2] Williamson (1986b), Carter and Manaster (1990), and Hauswald and Marquez (2006) consider similar cases of increasing returns to application scale, but not in relation to bank size.

Hence, bigger banks have better screening expertise and are more leveraged. This result is consistent with a trend of growing leverage in the banking industry: with small banks continuously edged out, the remaining banks become bigger and bigger, and, thus, the industry-wide leverage gets higher and higher.[3]

In this paper, financial intermediation arises naturally and coexists with direct finance. Screening expertise earns banks the informed funds' rate on the asset side, while they borrow from households at the uninformed funds' rate on the liability side. The difference in rates is the profit margin of financial intermediation. Besides investing in banks, households invest directly in projects that receive bank funding as the bank funding certifies their quality. In this paper, the allocation of households' funds between the intermediated and direct finance channels is uniquely determined. The financial intermediation affects the equilibrium outcome by expanding the pool of informed funds.[4]

The paper shows that a universal increase in banks' sizes may cause their asset qualities to decline. It is thus related to the literature that explains why loose lending standards are associated with lending booms; see Rajan (1994), Ruckers (2004), and Dell'Ariccia and Marquez (2006). A difference is that, in the present paper, an increase in size *drives* a fall in quality, whereas in that literature, both are driven by some other factor, such as bank managers' career concern in Rajan (1994), the distribution of borrowers' quality in Ruckers (2004), or the distribution of information in Dell'Ariccia and Marquez (2006).

In this paper, financial intermediation arises endogenously as banks want to enlarge their scales to exploit the increasing returns to scale. This is related to the literature that

[3]See Section 5 for a detailed discussion of empirical implications and the relevant empirical studies.

[4]In papers by Besanko and Kanatas (1993) and Holmstrom and Tirole (1997), entrepreneurs (firms) receive funds from both households and banks, as in the present paper. However, Besanko and Kanatas do not consider financial intermediation (i.e., banks drawing funds from households). While Holmstrom and Tirole include financial intermediation, they find that it does not affect the equilibrium outcome and that the allocation of funds between it and direct finance is indeterminate.

endogenizes financial intermediation with delegated monitoring and increasing returns to scale regarding the incentive costs of monitoring the bank.[5] The source of the increasing returns to scale is different: In this paper they are due to a general feature of expertise as mentioned above, while in that literature they are due to cross insurance. Moreover, direct finance and intermediated finance coexist in this paper, while not in that literature.

This paper finds bigger banks able to grow faster, implying a trend of increasing dominance, which is extensively examined in industrial economics; see Flaherty (1980), Gilbert and Newbery (1982), Budd, Harris and Vickers (1993), and, recently, Cabral (2011) and Besanko et al. (2011). Using the terminology of industrial economics, banks here engage in (differentiated product) Bertrand competition with limited capacity à la Kreps and Scheinkman (1983).

The rest of the paper is organized as follows. Section 1 sets up the basic model in which banks invest only their own funds. This model is analyzed in section 2 and extended in section 3 to encompass bank leverage. Section 4 lays out empirical implications, while section 5 discusses a modeling choice. Section 6 concludes. All proofs are relegated to the appendix.

1 Basic Model

The economy lasts for two dates (numbered 1 and 2) with no discounting and is populated by many small households, a continuum of $[0,1]$ of banks, and a continuum of $[0,1] \times [0,1]$ entrepreneurs. Banks are in perfect competition,[6] and each serves a continuum of entrepreneurs. All agents are risk neutral and protected by limited liability. Each entrepreneur has a project, while banks and households have funds.

[5] See Diamond (1984), Gale and Hellwig (1985), Williamson (1986a), Krasa and Villamil (1992), Winton (1995), and Cantillo (2004), among others, and see Gorton and Winton (2003) for a survey.

[6] The case of oligopolistic banks will be briefly discussed in section 6.

Funds are invested at date 1 either in projects or in a risk-free asset for which the gross rate of return is 1. A project requires an investment of $£B$ and may succeed, yielding $£Z$, or fail, yielding 0.[7] The probability of success is $\overline{q}$ for high-type projects and $\underline{q} < \overline{q}$ for low types. The fraction of high types is $\overline{n}$ and that of low types is $\underline{n} = 1 - \overline{n}$. We assume that

$$\overline{q}Z - B > 0 > (\overline{n}\overline{q} + \underline{n}\,\underline{q})Z - B. \tag{1}$$

That is, a high-type project has positive social value, but a randomly selected one does not.

Each bank $j \in [0, 1]$ has K_j units of funds, which captures the bank's size. The unit is so defined that out of 1 unit of funds, $£1$ can be invested in each of a continuum of mass 1 of projects. Without loss of generality, we assume K_j is a continuous, non-increasing function of j. In the basic model, banks are assumed to invest only their own funds, while bank borrowing will be incorporated in Section 3.

Households are small, but overall, their funds are so abundant that not all of these funds can be invested in entrepreneurs' projects. The remainder flows to the risk-free asset. It follows that households are satisfied with a gross rate of return 1.

Households know the prior distribution of the types of projects, but have no way of evaluating and screening them. However, banks can attain expertise in evaluating and screening projects. By assumption (1), only the high-type projects yield a surplus, so that the banks' screening services create a social value.

1.1 *Banks' Screening Expertise*

By spending $C(p)$, a bank acquires screening expertise of accuracy $p \in [0, 1]$. With the expertise of this accuracy, for each project it screens, the bank receives an independent

[7] If projects yield a return of $Z' \in (0, B)$ in the event of failure, then entrepreneurs can borrow up to Z' in a risk-free manner, which will not change the paper's results.

signal $\widetilde{s} = g(ood)$ or $b(ad)$ about the project's type according to

$$\Pr(\widetilde{s} = g|\widetilde{q} = \overline{q}) = 1, \quad \Pr(\widetilde{s} = b|\widetilde{q} = \underline{q}) = p, \quad \Pr(\widetilde{s} = g|\widetilde{q} = \underline{q}) = 1 - p, \tag{2}$$

where $\widetilde{q}$ is the true probability of success of the project. That is, high-type projects obtain a good evaluation with certainty, while low types receive a bad evaluation with probability p.[8] We call projects that obtain a good evaluation *good projects* and those that receive a bad evaluation *bad projects*.

To focus the analysis on banks, we assume that an entrepreneur does not know the type of his project before a bank evaluates it, and that he observes the result of the bank's evaluation. Households never observe this evaluation. Moreover, we assume that the screening accuracy of any bank is publicly observed, so the only information asymmetry between banks and entrepreneurs on the one hand, and households on the other, is over the evaluations of projects.

The cost function $C(\cdot)$ is convex over $[0, 1]$ and satisfies $C''(\cdot) > 0$, $C(0) = C'(0) = 0$ and $C'(1) = \infty$. Furthermore, $C'(p) = o(\frac{1}{(1-p)^2})$ around $p = 1$,[9] which ensures that the cost does not grow too fast in p.

Let $q_g(p)$ denote the posterior probability of success for a good project, $q_b(p)$ the posterior probability of success for a bad project, and $n_g(p)$ the probability of obtaining a good evaluation.[10] Then, $V_g(p) := q_g(p)Z - B$ is the social value of a good project; $S(p) := n_g(p)V_g(p)$ is the ex ante social surplus of a project if it is financed only when it obtains a good evaluation; and $d(p) := q_g(p)/q_b(p)$ measures the difference in quality

[8] Note that $C(p)$ is the cost of *obtaining screening expertise* of accuracy p, not that of evaluating a single project to this accuracy. That is, once a bank has paid $C(p)$, it can evaluates all projects at accuracy p. Allowing the accuracy of evaluating one project to depend on the resources spent for this particular evaluation would not qualitatively change the paper's results.

[9] That is, $C'(p)(1-p)^2 \to 0$ if $p \to 1$.

[10] With (2), $q_g(p) = \frac{\overline{n}\overline{q} + \underline{n}(1-p)\underline{q}}{\overline{n} + \underline{n}(1-p)}$, $q_b(p) = \underline{q}$, and $n_g(p) = \overline{n} + \underline{n}(1-p)$.

between good and bad projects. It is straightforward to show that

$$q_g'(p) > 0,\ d'(p) > 0,\ d(p) > 1 \text{ for } p > 0,\ S'(p) > 0,\ \text{and } S''(p) \leq 0\,. \tag{3}$$

These properties are all that is required for the analysis in this paper; the specific modeling of screening accuracy (2) is adopted for its simplicity.

Let $\underline{p}$ be the critical level of accuracy at which the social surplus of a good project just reaches 0, namely, $V_g(\underline{p}) = 0$. Then, $0 < \underline{p} < 1$ and a good project has a positive social value if and only if $p > \underline{p}$.[11] If the accuracy of evaluation is below $\underline{p}$, then even projects evaluated as being good are not financed and screening service of this accuracy is useless. A bank, if not choosing accuracy above $\underline{p}$, chooses $p = 0$. If a bank chooses so and thus to be uninformed, then it is identical to a household in making investments. We say that such a bank is *edged out* of the banking industry. If a bank opts to be informed by choosing some $p > \underline{p}$, we say that this bank stays in business and *survives competition*. We call funds provided by informed banks *informed funds* and funds provided by households (and uninformed banks) *uninformed funds*.

As for how banks can credibly communicate their evaluations of the projects to households, we make two assumptions.

Assumption 1: Banks' announcements of their evaluations of projects are not credible.

This assumption means that to certify good evaluations to households, banks must "put their money where their mouth is." A contract between an entrepreneur and a bank must involve the investment of the bank's funds. Such a contract is characterized by a pair (I, F), where I is the amount invested by the bank in the project and F is the *face rate of return*, or, simply, *the face rate*, of this investment so that $I \cdot F$ is the amount to

[11] This follows from the fact that $V_g'(p) = q_g' Z > 0$, $V_g(0) = (\overline{n}\overline{q} + \underline{n}\,\underline{q})Z - B < 0$, and $V_g(1) = \overline{q}Z - B > 0$ by the assumption in (1).

be repaid to the bank when the project succeeds; when it fails, by limited liability, no party gets anything.

Assumption 2: For a contract (I, F), the amount of investment, I, is observable to households, but the face rate, F, is not.

We justify this assumption on two grounds. First, in real life, the amount of investment is usually publicized or reported in the media, but the terms of the investment, which F represents, are not. Second, even if the terms are publicized, they are not a reliable guide to the actual rate of return, because the entrepreneur could easily sign another contract with the bank (for example, for consulting services) to arrange a side payment to the bank, and it would be too costly for households to check all the contracts between the two.

Because of these two assumptions, in order to credibly certify a project's quality, the bank needs to invest enough of its own funds, as will be shown. This ties banks' certification service, which is what they essentially sell to entrepreneurs, to the investment of their own funds. Then, bank size matters: the more funds a bank has, the more entrepreneurs it sells the certification service to, and, thus, the more profits it earns. This is how the paper endogenizes increasing returns to scale in the banking sector.

After an entrepreneur secures $£I$ of funds from an informed bank, he goes to the market for the funds of households. Households observe I, and based on it, infer the quality of the project. Projects that receive no bank funding are of lower-than-average quality, therefore do not attract household funding; their entrepreneurs withdraw from the market.

1.2 *Timing of Events and Information Structure*

The timing of events at date 1 is as below.

Stage 1: For each $j \in [0,1]$, bank j posts (p_j, R_j) – namely, the screening accuracy it attains and the expected rate of return it commits to charge for its funds.

Stage 2: Entrepreneurs each go to one informed bank and get their projects evaluated. The evaluation of a project is observed by the entrepreneur and the bank, but not by households. Based on the evaluations, the entrepreneurs submit a request for the bank's funds. If the bank has attracted too many entrepreneurs and the total demand for its funds is above its funding capacity, then rationing happens and only a fraction of the entrepreneurs get their requests satisfied. These entrepreneurs then sign a contract (I, F) with the bank, where I is observed by households. Afterwards, entrepreneurs seek funds from households. If they manage to acquire $£B$ altogether, they start their projects.

At date 2, the returns of the projects are realized and distributed to investors according to the contracts signed at date 1.

2 Size, Survival and Asset Quality

In order to find the subgame perfect equilibrium, we analyze the decisions in the following order. First, given (p, R) offered by a bank, the entrepreneurs coming to it decides their demand for the bank's funds, contingent on the evaluations of their projects. Second, given all banks' offers, $\{p_j, R_j\}_{j\in[0,1]}$, each entrepreneur decides to which bank he goes. Third, anticipating these decisions by entrepreneurs, banks decide on (p, R).

2.1 *Entrepreneurs' Decisions*

No entrepreneurs would come to an uninformed bank, whose funds serve no certification purposes. Consider the entrepreneurs who have come to an informed bank of size K that offers (p, R), with $p > \underline{p}$ and $R > 1$.[12] Then, the demand for the bank's funds by those

[12] As noted earlier, informed banks choose $p > \underline{p}$. Also, no informed bank posts $R \leq 1$, a return rate no higher than that of uninformed funds.

whose projects receive a good evaluation, henceforth called "*good entrepreneurs*", is

$$I(p,R) = \frac{V_g(p)}{Rd(p)-1}, \tag{4}$$

and the demand by *bad entrepreneurs,* namely those whose projects receive a bad evaluation, is 0.

This result is driven by two considerations. First, any demand of $I \geq I(p,R)$ of the bank's funds certifies the entrepreneur has received a good evaluation. Consider a bad entrepreneur who mimics a good one by demanding the same amount of bank funding. The bank, assessing the probability of success as q_b, charges him a face rate of R/q_b and demands a repayment of $I \cdot R/q_b$ when his project succeeds. If this investment I convinces households that the project is good, they are willing to finance the shortfall, $B-I$, at face rate $1/q_g$,[13] for a face value of $(B-I)/q_g$. If $V_m := Z - I \cdot R/q_b - (B-I)/q_g < 0$, the project's revenue in the case of success, Z, is insufficient to cover the liability outlay, $I \cdot R/q_b + (B-I)/q_g$. Then, the bank expects not to be fully repaid with $I \cdot R/q_b$, and therefore declines to lend I to the bad entrepreneur. Hence bank funding certifies only good evaluations if $V_m < 0$, or equivalently $I \geq I(p,R)$. Second, because bank funding is more costly than household funding, as $R > 1$, a good entrepreneur demands only the minimum bank funding necessary to certify the quality of his project – namely, $I(p,R)$ – and finances the shortfall with the cheaper household funding.

Consider now the aggregation of the individual demands for the bank's funds. Given an entrepreneur receives a good evaluation with probability $n_g(p)$ and the evaluations come independently, by the Law of Large Number, the total demand by one unit of entrepreneurs is $n_g(p)I(p,R)$. The bank can thus serve

$$\beta(p,R) = \frac{K}{n_g(p)I(p,R)} \tag{5}$$

units of entrepreneurs. If $M > \beta(p,R)$ units of entrepreneurs come to the bank, rationing

[13] So the expected rate of return is $q_g \cdot 1/q_g = 1$, which satisfies households.

happens, and each entrepreneur is served with probability $l = \beta(p, R)/M$. Otherwise, all the entrepreneurs are served by the bank. Hence, the probability of an entrepreneur being served is

$$l = \min(1, \frac{\beta(p, R)}{M}). \tag{6}$$

Conditional on being served, an entrepreneur obtains the difference of the social surplus of his project, $S(p)$, minus the profit surrendered to the bank. This profit equals $n_g(p)I(p, R)(R - 1)$ because the expected demand by him for the bank's funds is $n_g(p)I(p, R)$ and each pound of them earns the bank a net profit of $R-1$. With $I(p, R)$ given by (4) and $S(p) = n_g(p)V_g(p)$, the expected payoff of a served entrepreneur is

$$\Pi(p, R) = S(p)\frac{(d(p) - 1)R}{d(p)R - 1}. \tag{7}$$

Now, consider entrepreneurs' decisions on which banks to borrow from, given all banks' deals, $\{p_j, R_j\}_{j\in[0,1]}$. All the banks that attract entrepreneurs to come offer an incoming entrepreneur the same expected payoff, $\widehat{\Pi}$, because no entrepreneurs would go to a bank that offers less if they can get $\widehat{\Pi}$ from some other banks. Therefore, if a bank offering (p, R) attracts $M > 0$ units of entrepreneurs, then

$$l \cdot \Pi(p, R) = \widehat{\Pi}, \tag{8}$$

where l is given by (6). The allocation of entrepreneurs to banks, $\{M_j\}_{j\in[0,1]}$, and $\widehat{\Pi}$ are determined by (8) and the following market-clearing condition:

$$\int_{j\in[0,1]} M_j = 1. \tag{9}$$

2.2 *Banks' Decisions: Increasing Returns to Scale and Competition*

Having examined entrepreneurs' decisions, we move on to find the best response of a bank given the choices of (p, R) by all the other banks. Given there is a continuum of banks,

each bank has only a negligible effect on $\widehat{\Pi}$, so each bank takes $\widehat{\Pi}$ as given when choosing (p, R). Given $\widehat{\Pi}$, a bank can choose to be uninformed and thereby get 0 economic profit. If it chooses to be informed, it has to ensure that its offer, (p, R), can attract entrepreneurs, i.e., that $\Pi(p, R) \geq \widehat{\Pi}$ by (8). The bank has no incentives to make $\Pi(p, R) > \widehat{\Pi}$ and induce excess demand.[14] Therefore, at the optimum, $\Pi(p, R) = \widehat{\Pi}$, which, together with (7), implies that after attaining accuracy p, the bank charges the following interest rate:

$$R(p; \widehat{\Pi}) = \frac{\widehat{\Pi}}{d(p)\widehat{\Pi} - (d(p) - 1)S(p)}. \tag{10}$$

Neither does the bank induce under-demand.[15] It follows that if a bank of size K chooses to be informed at accuracy p, all its funds are lent out at return rate $R(p; \widehat{\Pi})$, and its value is

$$K \cdot (R(p; \widehat{\Pi}) - 1) - C(p). \tag{11}$$

The bank's decision problem is to find a $p \in [0, 1]$ to maximize this value, taking $\widehat{\Pi}$ as given. Let $\rho(K, \widehat{\Pi})$ be the optimal p and $\Theta(K, \widehat{\Pi})$ the optimal value of this decision problem. Then, the bank chooses to be informed only if $\Theta(K, \widehat{\Pi}) \geq 0$, and if it does so, it picks

$$p = \rho(K, \widehat{\Pi}) \tag{12}$$

$$R = R(\rho(K, \widehat{\Pi}), \widehat{\Pi}), \tag{13}$$

where $R(p, \widehat{\Pi})$ is given by (10). Since the bank induces neither excess-demand nor under-demand, it receives the following number (units) of entrepreneurs and serves them all:

$$\beta(K, \widehat{\Pi}) := \beta(\rho(K, \widehat{\Pi}), R(\rho(K, \widehat{\Pi}), \widehat{\Pi})), \tag{14}$$

where $\beta(p, R)$ is given by (5).

[14] Otherwise, it would increase the interest rate, R, which decreases $\Pi(p, R)$, and lend all its funds out at this increased rate.

[15] If the bank has only part of its funds earn return rate $R(p; \widehat{\Pi})$, it would lower the rate a little to $R - \varepsilon$, which induces over-demand, as $\Pi(p, R - \varepsilon) > \widehat{\Pi}$, so all its funds earn return rate $R - \varepsilon$.

The lemma below establishes a lower bound for the equilibrium payoff of entrepreneurs.

Lemma 1 *If* $\widehat{\Pi} \leq \frac{d(1)-1}{d(1)} S(1)$, *then* $\Theta(K, \widehat{\Pi}) = \infty$ *for any* $K > 0$.

The lower bound for $\widehat{\Pi}$ exists because of competition between banks. If $\widehat{\Pi}$ were below the bound, then a particular bank would undercut all the other banks and obtain a big profit by providing a screening service of such accuracy p that the payoff to entrepreneurs, $\Pi(p, R)$, is above $\widehat{\Pi}$, namely, what they can obtain elsewhere, for even $R \to \infty$. By this lemma, any function of $\widehat{\Pi}$, such as $\Theta(K, \widehat{\Pi})$, is meaningfully defined only for for $\widehat{\Pi} > (d(1) - 1)S(1)/d(1)$, which, therefore, is a precondition for any proposition below where $\widehat{\Pi}$ is taken as given.

Now we can explain the two forces that shape the economic implications of bank size: increasing returns to scale and competition. Each force presents itself at both profit and marginal profit levels. These two forces, at these two levels, are formalized as follows.

Proposition 1 *For* $(K, \widehat{\Pi})$ *at which* $\infty > \Theta(K, \widehat{\Pi}) > 0$ *(namely, the bank chooses to be informed),*

(i) $\partial\Theta/\partial K > 0$, $\partial^2\Theta/\partial K^2 > 0$, *and* $\partial\rho/\partial K > 0$;

(ii) $\partial\Theta/\partial\widehat{\Pi} < 0$, *and* $\partial\rho/\partial\widehat{\Pi} < 0$.

Result (i) is concerned with *increasing returns to scale* at the two levels. First, at the level of bank profit, we have $\partial\Theta/\partial K > 0$, that is, bigger banks get higher value from being informed. A bigger capacity benefits informed banks by raising both extensive margin and profit (or intensive) margin. A bank becomes informed only if doing so enables it to charge $R > 1$ for its funds; thus, the more funds a bank deploys, the higher profit it earns on becoming informed. Also, the bigger the bank, the higher the profit margin, because

by (10), R increases with the screening accuracy,[16] which, as $\partial\rho/\partial K > 0$, increases with size.

Second, increasing returns to scale at the level of marginal profit drive $\partial\rho/\partial K > 0$, that is, bigger banks choose higher screening accuracy. The choice of screening accuracy, $\rho(K, \widehat{\Pi})$, which maximizes (11), satisfies the following first order condition for p:

$$K\widehat{\Pi}\frac{(d(p)-1)S'(p) + d'(p)(S(p)-\widehat{\Pi})}{[d(p)\widehat{\Pi} - (d(p)-1)S(p)]^2} = C'(p). \tag{15}$$

The marginal profit from higher accuracy, which appears on the left-hand side of the equation, is proportional to the size, K. Intuitively, screening expertise of higher accuracy enables the bank to charge a higher interest rate, which augments the bank's profit farther if the bank is larger.

From $\partial\rho/\partial K > 0$ it follows that the larger the bank, the bigger the marginal value of size (i.e. $\partial\Theta/\partial K$), because by the envelope theorem $\partial\Theta/\partial K = R(\rho(K, \widehat{\Pi}), \widehat{\Pi}) - 1$ and thus $\partial^2\Theta/\partial K^2 = R'_p \cdot \partial\rho/\partial K > 0$.

Result (ii) is concerned with two effects of competition, where competition is represented by $\widehat{\Pi}$, the payoff a bank has to give entrepreneurs for attracting them to come. First, at the level of profit, we have $\partial\Theta/\partial\widehat{\Pi} < 0$; that is, fiercer competition lowers banks' profit. This is because it forces banks to surrender more payoff to entrepreneurs and, thus, to obtain less from the certification service.

Second, we have $\partial\rho/\partial\widehat{\Pi} < 0$, that is, fiercer competition drives banks to lower screening accuracy. It does so by squeezing the marginal profit to banks from an accuracy increment, which both widens the profit margin and enlarges the business scale. Screening of higher accuracy increases the social value of a project, $S(p)$, and thereby widens the profit margin from each project screened, $S(p) - \widehat{\Pi}$. The total marginal profit so generated

[16] Note that $-[d(p)\widehat{\Pi} - (d(p)-1)S(p)]'_p = (d-1)S' + d'(S-\widehat{\Pi})$, which is positive because $d > 1$, $d' > 0$ and $S' > 0$ by (3), and $S - \widehat{\Pi}$, the surplus the bank gets from each project evaluated, is positive if the bank chooses to be informed.

is proportional to the number of projects screened, which decreases with $\widehat{\Pi}$: the higher the payoff to the entrepreneurs ($\widehat{\Pi}$), the lower the rate charged (R); and the more the bank's funds demanded by each good entrepreneur for certification ($I'_R < 0$ by 4), and the fewer the entrepreneurs served. Also, higher accuracy enables the bank to serve more entrepreneurs because entrepreneurs, when more accurately evaluated as being good, need less of the bank's funds for certification ($I'_p < 0$ by 4). The gain to the bank from the added entrepreneurs is proportional to the profit margin, $S(p) - \widehat{\Pi}$, which obviously decreases with $\widehat{\Pi}$. Therefore, a higher $\widehat{\Pi}$ diminishes the marginal profit of an accuracy increment to banks in both dimensions.

The increasing returns to scale imply that only big banks survive competition, as stated in the following proposition.

Proposition 2 *Let* $\underline{K}(\widehat{\Pi})$ *be the largest root of* $\Theta(K, \widehat{\Pi}) = 0$.

(i): $\underline{K}(\widehat{\Pi})$ *exists and is strictly positive, and* $\Theta(K, \widehat{\Pi}) > 0$ *if and only if* $K > \underline{K}$.

(ii): $\underline{K}'(\widehat{\Pi}) > 0$.

Result (i) says that a bank survives competition (i.e., $\Theta \geq 0$) only if its size is above a threshold (i.e., $K \geq \underline{K}$), while smaller banks are edged out. Result (ii) says that when competition becomes fiercer (i.e., a larger $\widehat{\Pi}$), the threshold of survival rises. This is because the more payoff needed to give entrepreneurs, the less a bank gets from serving one entrepreneur on becoming informed; in order to cover the cost of attaining screening expertise, therefore, the more entrepreneurs it needs to serve, which requires a larger funding capacity.

The two results together suggest that the banking industry is subject to a trend of increasing concentration.[17] The pressure to survive and the motive to exploit the increasing returns to scale drive banks into a race for expansion, which raises $\widehat{\Pi}$ (as we

[17] For supportive empirical evidence, see the empirical studies cited in footnote 1.

will show), so that all banks face even fiercer competition. Moreover, if the extent of expansion is proportion to the marginal value of size, namely to $\partial\Theta/\partial K$, then result $\partial^2\Theta/\partial K^2 > 0$ (Proposition 1.i) suggests that smaller banks are able to expand less, thus losing out in the race. Altogether, smaller banks are continuously edged out, leaving the banking industry more and more concentrated.

Having examined the decisions of entrepreneurs and banks, we now proceed to define the equilibrium formally and examine its existence and properties.

2.3 *Equilibrium*

Banks choose to be informed if the value from doing so, Θ, is nonnegative, which, by Proposition 2, is the case if and only if $K \geq \underline{K}(\widehat{\Pi})$. As we assume K_j to be non-increasing in j, it follows that there is a threshold $t \in [0, 1]$, such that bank j chooses to be informed if and only if $j \leq t$. There are two cases for the value of t. One, $t = 1$, namely, all banks become informed; in this case $K_1 \geq \underline{K}(\widehat{\Pi})$. The other, $t < 1$, namely, smaller banks are edged out; in this case, $K_t = \underline{K}(\widehat{\Pi})$, because K_j is assumed continuous in j. Note that in the latter case, there may be $a, b \in (0, 1)$ such that $a < t < b$ and $K_a = K_b = \underline{K}$, that is, the marginal banks, which are of size $\underline{K}$, play a mixed strategy, some of them being informed, the rest uninformed.

As banks' decisions on whether to be informed are summarized by variable t and each of them takes $\widehat{\Pi}$ as given, an equilibrium is thus represented by a pair of $(t, \widehat{\Pi})$, defined as follows.

Definition 1 *A pair of* $(t, \widehat{\Pi})$ *is an equilibrium if*

(i) Given $\widehat{\Pi}$,

$$K_t = \underline{K}(\widehat{\Pi}) \text{ if } t < 1 \text{ or } K_1 \geq \underline{K}(\widehat{\Pi}) \text{ if } t = 1; \tag{16}$$

(ii) the market clears:

$$\int_0^t \beta(K_j, \widehat{\Pi})dj = 1.^{18} \tag{17}$$

Condition (17) is derived from (9) and the fact that no banks induce excess demand or under-demand. Once $(t, \widehat{\Pi})$ is pinned down, the equilibrium decisions of banks and entrepreneurs follow straightforwardly. Bank j chooses to be informed if and only if $j \leq t$; informed banks choose (p, R) as given by (12) and (13); $\beta(K, \widehat{\Pi})$ (given by 14) units of entrepreneurs go to an informed bank of size K and demand $I(p, R)$ (given by 4) of its funding when their projects receive a good evaluation and demand nothing otherwise. The equilibrium market share of bank j is thus:

$$\widehat{\beta}_j = \begin{Bmatrix} \beta(K_j, \widehat{\Pi}) \text{ if } j \leq t \\ 0 \text{ if } j > t \end{Bmatrix}.$$

Proposition 3 *A unique equilibrium exists and has the following properties.*

(i) If a positive measure of informed banks increase their sizes, then the payoff of entrepreneurs $(\widehat{\Pi})$ *increases, and all the other informed banks lower their screening accuracies and interest rates.*

(ii) If all banks increase size without changing their market shares, $\{\widehat{\beta}_j\}_{j\in[0,1]}$*, then they all lower screening accuracy and interest rate.*

Result (i) says that expansion by some banks toughens competition (i.e., $\widehat{\Pi}$ higher) and weakens their competitors, who consequently lower the quality of their screening expertise and the price of their funds. The expanded banks face the same negative effects of competition, but are blessed by the increasing returns to scale, which enable them to choose greater screening accuracy (i.e., $\partial\rho/\partial K > 0$) and thereby charge a higher price. For an enlarged bank, these positive effects dominates the negative effects if it expands

[18] As will be shown in the proof of Proposition 3, $\partial\beta/\partial K > 0$, that is, $\beta(K_j, \widehat{\Pi})$ is non-increasing with j. Thus as a function of j, $\beta(K_j, \widehat{\Pi})$ is integrable.

far more than all the other expanding banks, in which case its expansion not only weakens the competitors, but also strengthens itself. Therefore, not only do banks want to expand, but also they want to expand far more than all the others.

Result (ii) gives one condition under which this race for expansion weakens all the banks: it adds no market share to any bank. For an intuition, note that expansion always toughens competition, which tends to decrease the profit margin of providing certification service for all banks. If this decrease in profit margin is not compensated by an increase in business scale – namely, if banks cannot sell certification service to more entrepreneurs – banks gain less from screening expertise, and therefore attain the less of it.

With all banks' screening expertise weakened, the default risks of their assets rise.[19] Note that banks still invest only in projects evaluated as good, but with the evaluations becoming less accurate, the composition of these projects gets worse: a smaller fraction of them are high types, a bigger one low types.

2.4 *Welfare Properties of the Equilibrium*

The equilibrium is ex post efficient, because all the good projects are financed and none of the bad ones is. As for ex ante efficiency, we define the first-best allocation as the choice of the social planner if she can allocate entrepreneurs to banks and pick a level of accuracy for each bank, and the second-best allocation as the planner's choice if she has to respect the equilibrium market shares of banks, but picks screening accuracy for each informed bank. The first best allocation is more efficient than the second best one, which, we show below, is more efficient than the equilibrium allocation.

Suppose the planner chooses accuracy p for bank j. Then the bank's service generates a social value of $S(p)$ from each of the $\widehat{\beta}_j$ units of entrepreneurs whom it serves in equi-

[19] Mathematically, the default probability of the projects bank j invests is $1 - q_g(p_j)$, which increases when p_j decreases, because $q_g'(\cdot) > 0$.

librium. Overall, thus, the bank generates a social value of $\widehat{\beta}_j \cdot S(p)$, while the social cost of attaining the accuracy is $C(p)$. The social planner's problem for bank j is therefore:

$$\max_{0 \leq p \leq 1} \widehat{\beta}_j S(p) - C(p).$$

The second-best choice of quality, denoted by p_j^*, satisfies the following first order condition:

$$\widehat{\beta}_j S'(p^*) = C'(p^*). \tag{18}$$

Proposition 4 *For any informed bank j, $\widehat{p}_j > p_j^*$. That is, compared to the second-best allocation, banks overspend on screening expertise.*

For an intuition, refer back to the discussion of why $\partial \rho / \partial \widehat{\Pi} < 0$ (see Proposition 1.ii). There we show that for a bank, higher screening quality generates two benefits, one from more entrepreneurs to be served, the other from a widened profit margin $S(p) - \widehat{\Pi}$. As $(S(p) - \widehat{\Pi})' = S'(p)$, the latter benefit accrues equally to the social planner. However, the former benefit does not accrue to the social planner because the planner takes as given the number of entrepreneurs allocated to each bank. It is the motive of attracting more entrepreneur with higher screening quality that drives banks to overspend on screening expertise.

Essential to the proposition is the model's feature that banks' funds serve mainly certification purposes and the investments of the projects are mainly financed by households. Should the household sector be absent, an entrepreneur's demand for the bank's funds would be fixed at B, the investment need. The bank would serve K/B units of entrepreneurs, independent of its screening quality. Improved screening quality would not bring more entrepreneurs to the bank, therefore, the overspending result would not arise.

For an implication of this result, consider where the resources are spent to attain or improve screening expertise. A fraction of the resources might be spent on IT infrastructure. But for the banking sector, probably a bigger fraction is spent in attracting human

capital with high salaries and/or bonuses. For example, over 2005–2010, the average ratio of compensation and benefits to overall non-interest expenses was 65% for Goldman Sachs and 64% for Morgan Stanley,[20] while, in contrast, this ratio for U.S. manufacturing sector is 11%.[21] Moreover, from an economics point of view, these payments of compensation and benefits may work more as a fixed cost than as a marginal cost because it seems that very often they are outlaid independently of the banks' performance rather than anchored to it.[22] If we accept that a big fraction of $C(p)$ is thus spent, then the proposition suggests that the banking industry indeed hands out excessive payments, and a policy to cap them could improve ex ante efficiency.

The next section extends the model to encompass bank leverage, whereby we show that there is a complementarity between leverage and screening quality.

3 Bank Leverage

In the analysis thus far, banks do not borrow from households; rather, they invest only their own funds. In this section, we assume that before banks choose screening accuracy, they can borrow funds from households, while using their own funds as the equity.[23] The investing households, as debt-holders, are repaid prior to the banks, the equity holders.

Banks' advantage over households in screening expertise drives banks to borrow. It

[20] Calculated from the annual reports of the two banks published on their websites.

[21] See "The misery of manufacturing," The *Economist*, pages 75-76, 27/09 - 03/10, 2003.

[22] For example, for year 2008, Citigroup and Merrill Lynch both paid their employees each a bonus of one million or more dollars, although the banks suffered a huge loss; see "Million-dollar bonus breakdown to reignite US bank controversy" (*Financial Times*, 31/07/2009). Banks often resort to the need to retain key human capital, rather than that to provide incentives, when coming to justify high bonuses.

[23] The paper assumes banks do not issue outside equities to households, possibly due to some friction of costly state verification in the manner of Townsend (1979), Diamond (1984), and Gale and Hellwig (1985).

enables them to earn the rate of informed funds on the asset side, while they repay the rate of uninformed funds to households on the liability side, the gap between these two rates producing the profit margin of borrowing. Individual banks take this profit margin as given, and so long as it is positive, they want to borrow as much as possible. To limit their borrowing, we introduce risk-shifting problems in the manner of Jensen and Meckling (1976). This requires the risks of the projects to be correlated,[24] while the analysis so far is independent of such correlation. To simplify the exposition, we assume that the risks of projects are perfectly correlated. Specifically, foreseen at date 1, at date 2 the economy is in one of three possible states, $\{\phi, 1, 2\}$, occurring with probability $1-\overline{q}$, $\overline{q}-\underline{q}$, and $\underline{q}$, respectively. In state ϕ, no projects succeed; in state 1, only high-type projects succeed and low types fail; and in state 2, both types succeed. So high-type projects succeed in both states 1 and 2, thus with probability $\overline{q}$, and low types succeed in state 2 only, thus with probability $\underline{q}$.

3.1 *The Risk-Shifting Problem and Leverage Ratio*

Suppose a bank borrows D units of funds from households at face rate f, so that its liability is Df. The risk-shifting problem of the bank is that if D is too large, the bank may want to invest in bad projects at a lower expected return rate but a higher face rate than it obtains by investing in good projects.

Let F be the face rate of investing in good projects, that is, $F = R/q_g$, and F' be that of investing in bad projects, which succeed with probability q_b. The value of F' lies between F and $F \cdot q_g/q_b$. On the one hand, the bank rejects any face rate below F. On the other hand, bad entrepreneurs cannot afford a face rate above $F \cdot q_g/q_b$, namely an expected rate of return above R, by the argument leading to (4). Then, for some

[24] Otherwise, as each bank finances a continuum of projects, the risks on its asset side will be completely diversified away and no risk-shifting problems will arise.

$\alpha \in (0,1)$,

$$F' = ((1-\alpha)\frac{q_g}{q_b} + \alpha)F. \tag{19}$$

Note that $q_b F' < q_g F$, but $F' > F$, that is, weighted against the risk, the investment in bad projects is worse than that in good projects, but contingent on success, the former delivers a higher return than the latter does. This hallmarks a typical circumstance liable to risk-shifting problems. To prevent them, the leverage ratio, $L := D/K$, should be capped, as the following lemma shows.

Lemma 2 *If a bank with K units of equity funds borrows D at face rate f, attains accuracy p, and charges return rate R, then it does not invest in bad projects if and only if the leverage ratio L satisfies:*

$$L \leq \frac{\alpha(1-\frac{1}{d(p)})R}{(\overline{q}-\underline{q})f - (1-\frac{1}{d(p)})R}. \tag{20}$$

If the inequality holds, the bank repays its debt with probability $\overline{q}$ and

$$f = 1/\overline{q}. \tag{21}$$

The bank repays its debt with probability $\overline{q}$ because it does so whenever high-type projects succeed, which is in turn because the debt claims are senior to the equity claim held by the bank. The debt holders earn an expected rate of return of $\overline{q}f$ and are satisfied with a return rate of 1. Hence (21) holds.

3.2 *Complementarity between Leverage and Screening Accuracy*

In this subsection we find the equilibrium leverage of banks and show a complementarity between leverage and screening accuracy. Banks, taking the profit margin of borrowing as given, want to borrow as much as they can fend off the risk-shifting problems. They are thus leveraged to the upper bound given by (20). With f given by (21) and R as a

function of p given by (10), bank j's leverage ratio, L_j, is

$$L_j(p_j) = \frac{\alpha(1 - \frac{1}{d(p_j)})R(p_j, \widehat{\Pi})}{(\overline{q} - \underline{q})/\overline{q} - (1 - \frac{1}{d(p_j)})R(p_j, \widehat{\Pi})}. \tag{22}$$

This equation describes how the leverage ratio of a bank depends on its screening accuracy.[25]

Furthermore, leverage feeds back to screening accuracy. If bank j is leveraged at ratio L_j, then its screening accuracy is given by

$$p_j(L_j) = \rho(K_j(1 + L_j), \widehat{\Pi}). \tag{23}$$

That is, in determining the choice of screening accuracy, debt-financed funds, $D = K \cdot L$, play an equal role as the equity funds, K. This is because borrowed funds and equity funds contribute in equal terms to the marginal value of higher accuracy, for which an intuition is as follows. By the preceding lemma, the probability of debt being repaid, fixed at $\overline{q}$, is independent of the accuracy choice, p. Hence, so is the marginal cost of debt (D), as is the marginal cost of equity (K). Moreover, debt and equity contribute in equal terms to a bank's revenue at given p (i.e. $(K + D)R(p)$) and thus to its marginal revenue (i.e. $(K + D)R'_p$). Therefore, D and K contribute in equal terms to both the marginal benefit and the marginal cost of higher p.

Equations (22) and (23) together yield the following.

Proposition 5 *There is a complementarity between screening accuracy and leverage: for any bank j, $L'_j(p_j) > 0$ and $p'_j(L_j) > 0$. That is, on the one hand, a bank with more-accurate screening expertise is leveraged at a higher ratio; on the other hand, higher leverage leads to greater accuracy.*

[25] Or, rather the rational expectation of the accuracy, as the leverage choice is decided before the accuracy choice.

Intuitively, that $L_j'(p_j) > 0$ because the greater the screening accuracy p, the starker the difference in quality between good and bad projects (i.e. d) and the higher the rate charged (because $R_p' > 0$). Both lead to bigger value destruction by risk shifting (as $R - q_b F' = \alpha(1 - 1/d)R$ by 19), thus inducing smaller incentives to do that, which allows for higher leverage. That $p_j'(L_j) > 0$ because higher leverage augments the bank's size farther, which, due to the increasing returns to scale at the level of marginal profit, induces the bank to attain greater accuracy.

The next section presents empirical implications of the paper.

4 Empirical Implications

A. *Bigger banks are leveraged at higher ratios*: the bigger the bank, the higher the screening accuracy due to increasing returns to scale, and, by the complementarity, the higher the leverage.[26] We show in Proposition 1 $\partial^2\Theta/\partial K^2 > 0$, which suggests bigger banks able to enlarge their equity (i.e. K) farther if the extent of equity enlargement increases with the marginal value of equity, $\partial\Theta/\partial K$. Together with implication A, therefore, the paper suggests that *bigger banks can expand more, both by enlarging equity farther and by being leveraged higher*. This suggests a trend of increasing dominance in the banking industry.

Implication A is consistent with many empirical findings. Liang and Rhoades (1991), with a sample of 4751 US banking firms over 1979–86, report on Table II that the total asset (TA) negatively and significantly affects the equity/asset ratio (E/A). Akhavein et al. (1997), analyzing big U.S. banks over 1980–1990, find that after merger and acquisitions (M&A), consolidated banks widen the negative difference in E/A from their peer banks by 6 basis points. Demsetz and Strahan (1997), examining large U.S. bank holding companies (BHC) over 1980–1993, document in Table 3 a strong, significant, and negative correlation

[26] Mathematically, if $A_i := K_i + D_i > A_j := K_j + D_j$, then by (23) and $\partial\rho/\partial K > 0$, $p_i > p_j$, and hence $L_i > L_j$ since $L'(p) > 0$.

between size and E/A. This strong negative correlation is also found, more recently, by Lepetit et al. (2008) for a set of European banks over 1996–2002 and by Haq and Heaney (2012) for 117 European financial institutions over 1996–2010.

While this empirical literature often attributes the higher leverage of bigger banks to the benefits of diversification, the present paper finds it may come, orthogonal to diversification, from the complementarity between leverage and screening expertise. The two arguments diverge in the link between the size of a bank and the quality of its individual loans. No link is implied by the diversification argument, while this paper implies the following.

B. *Obtaining funding from bigger banks certifies that the borrowing firms are of higher quality.* Bigger banks have more-accurate screening expertise. Thus the projects that they evaluate as being good and then invest in are more likely to be among the high types, therefore of higher quality.[27]

For this result, direct empirical support is provided by Ross (2010). He documents that loans from three dominant banks (J.P. Morgan Chase, Bank of America, and Citigroup, accounting for more than 55% of the U.S. commercial loan market) over 2000–2003 induced their borrowers' stock prices to jump higher, were issued at lower interest rates, and were "less likely to be protected by a borrowing base," altogether suggesting that these banks "provide a higher level of certification." (both quotations on p. 2731). Also, Hao (2003) documents, using a sample of U.S. banks over 1988–1999, an inverse link between bank size and loan yield spread, which, the author suggests, may be explained by bigger banks picking borrowers of higher credit-quality. Indirectly, Billett et al. (1995) report, using a sample of corporate loans over 1980–1989, that greater abnormal returns of the borrowers' shares are associated with loans from banks with higher credit ratings,[28]

[27] Mathematically, if $A_i := K_i + D_i > A_j := K_j + D_j$, then by (23) and $\partial\rho/\partial K > 0$, $p_i > p_j$. Hence $q_g(p_i) > q_g(p_j)$.

[28] For other empirical studies on how borrowers' stock prices respond to news of obtaining or renewing

which Poon et al. (2009) show is strongly and positively correlated with bank size.[29]

C. *The banking industry displays a trend of growing leverage.* This is because small banks keep being edged out and only big banks remain, which, by implication A, are more leveraged.

This trend is empirically well documented; see, among others, Berger et al. (1995; figure 1) for the U.S. over 1840–1990, Saunders and Wilson (1995; figures 4–6) for the UK, Canada and U.S. over 1893–1991, Hortlund (2005; figure 2) for Sweden over 1870–2001, and Miles et al. (2012; figure 1) for the UK over 1880–2010.

D. *An industry-wide rise in leverage tends to increase concentration in the banking industry.* This is because it enlarges the capacity of all banks and thereby intensifies competition between them, consequently edging out more small banks.

Berger, Demsetz and Strahan (1999) document that an important factor contributing to the substantial consolidation in U.S. banking industry over 1988–1997 (when the number of banks fell by 30%) was the improvement in financial conditions, such as low interest rates, which made it more profitable for banks to increase leverage. Moreover, they find it puzzling that "M&A activity in banking appears to respond more to low interest rates ... than does M&A activity in non-financial industries, despite the fact that stock deals are more common than cash acquisitions in banking ..." (p. 149). This paper gives an account for this phenomenon by showing that a larger capacity delivers greater benefits in the financial sector than it does in a non-financial sector. In the former, we show in the discussion of Proposition 1(i) that a larger capacity benefits a bank in both

bank loans, see, among others, Mikkelson and Partch (1986), James (1987), Lummer and McConnell (1989), and Best and Zhang (1993).

[29] Moreover, Demsetz and Strahan (1997) find that larger BHC take a greater proportion of commercial and industrial loans relative to securities, and Akhavein et al. (1997) find that after M&A, the consolidated banks shift from securities to loans. As loans demand more screening expertise than securities to invest, both papers suggest bigger banks have a higher level of screening expertise.

extensive margin and profit margin, whereas in the latter, it usually delivers no benefit in profit margin. Lastly, Berger, Kashyap and Scalise (1995) find that deregulation of deposit ceiling rates contributed to the consolidation of the U.S. banking industry over 1979–1994. That is consistent with the paper's findings, because the deregulation allowed banks to absorb more deposits, which increased leverage.

An industry-wide rise in leverage could be induced by a regulatory loophole; for example, the off-balance investment vehicles to the Basel II accords.[30] The thus-created "shadow banking system" contributed substantially to the massive increase in bank leverage over the decade leading up to the 2008 crisis.

5 *Oligopolistic Banks*

In this paper, banks are modeled in perfect competition in the sense that each bank takes entrepreneurs' equilibrium payoff, $\widehat{\Pi}$, as given. This approach offers a simple way to disentangle the two forces that this paper identifies as important in shaping the economic implications of bank size, namely, increasing returns to scale and competition. In many real-life economies, however, the banking industry is an oligopoly. This section briefly discusses how the main results of this paper apply under such circumstances.

To simplify the exposition, consider a two-bank case, where banks 1 and 2 have funds K_1 and K_2 respectively and compete for one unit of entrepreneurs; all the other aspects are as they were modeled in section 1. Following the analysis of subsection 2.1, after banks have chosen $(p_j, R_j)_{j=1,2}$, the allocation of entrepreneurs to banks, (M_1, M_2), and

[30] By Basel II, the requirement of capital buffer is based on the risk-weighted assets. Off-balance investment vehicles shift the risky assets off the bank's books, therefore circumvent the need to hold capital for the assets, which allows an increase in leverage.

the payoff of entrepreneurs, $\widehat{\Pi}$, are determined by the following three equations:

$$l_1\Pi(p_1, R_1) = l_2\Pi(p_2, R_2) = \widehat{\Pi} \tag{24}$$

$$M_1 + M_2 = 1, \tag{25}$$

where

$$l_j = \max(1, \frac{K_j}{M_j n_g(p_j) I(p_j, R_j)}). \tag{26}$$

Let the solution for M_j be $M_j(p_j, R_j; p_{-j}, R_{-j})$ for $j = 1, 2$.

Now consider the decisions of the two banks. As in the case of perfect competition, it is not optimal for a bank to induce over-demand. However, unlike the previous case, a bank can now be too large, in the sense that it is optimal for it to induce under-demand. The paper assumes this case away because it hardly captures a real-life circumstance. Then, given the other bank's choice (p_{-j}, R_{-j}), bank j chooses (p_j, R_j) such that

$$M_j(p_j, R_j; p_{-j}, R_{-j}) = \frac{K_1}{n_g(p_j) I(p_j, R_j)}, \tag{27}$$

and its best response is to be found by solving the following problem:

$$\max_{p_j, R_j} K_j(R_j - 1) - C(p_j), \text{s.t. } (27). \tag{28}$$

The best responses pin down the subgame perfect equilibrium.

In this setting of oligopolistic banks, as in that of perfect competition, there exist the forces of increasing returns to scale and competition, as formally presented in Proposition 1. Because $l_j = 1$ for $j = 1, 2$, it follows from (24) that $R_j = R(p_j, \widehat{\Pi})$, with $R(p, \widehat{\Pi})$ given by (10). Thus the objective in problem (28) is the same as that in the case of perfect competition, given by (11). Therefore, if $\widehat{\Pi}$ could be taken as a given to a bank, then Proposition 1 would hold here also. However, here $\widehat{\Pi}$ cannot be taken as given when the size of one bank is changed, because now any bank has a non-neglegible effect on $\widehat{\Pi}$. Therefore, in this setting of oligopolistic banks, the effects of increasing returns to

scale and those of competition cannot be disentangled. However, some results parallel to Proposition 3 can be derived, as follows.

Proposition 6 *Assume the two banks are of such sizes that they both choose to be informed and not to induce under-demand. The subgame perfect equilibrium uniquely exists and has the following properties.*

(i) If a bank increases its size, then the payoff of entrepreneurs $(\widehat{\Pi})$ *increases and the other bank lowers its screening accuracy and interest rate.*

(ii) If both banks increase size without changing their market shares, then they both lower screening accuracy and interest rate.

The proposition can be proved in the same way in which Proposition 3 is proved.

6 Conclusion

In a framework where banks can attain expertise to screen projects, while the general public (namely households) cannot, this paper examines the causal effects of bank size for banks' survival, asset quality and leverage. The paper finds the following.

First, banks' screening expertise generates increasing returns to scale, which help bigger banks survive competition better, and drive banks into a race for expansion. This race intensifies competition between banks and thereby dampens their incentives to attain or improve screen expertise. Moreover, in this race, small banks tend to lose out, subjecting the banking industry to a trend of increasing concentration.

Second, if all banks expand without much changing their market shares, then as a consequence of fiercer competition, their screening expertise all declines and asset quality all falls, in spite of the increasing returns to scale.

Third, there is a complementarity between a bank's leverage and the level of its screening expertise.

Fourth, there is a sense in which banks overspend on screening expertise. If, as seems plausible, a big fraction of this spending is used to attract human capital with high bonuses or salaries, then this overspending result suggests that the banking industry indeed hands out excessive payments.

Appendix: Proofs

For Lemma 1:

Proof. There are two cases. First, we show that if $\widehat{\Pi} < \frac{d(1)-1}{d(1)}S(1)$, a bank gets an infinitely large profit, namely, $\Theta(K,\widehat{\Pi}) = \infty$. Note first that $\frac{d(\cdot)-1}{d(\cdot)}S(\cdot)$ is an increasing function because both $d'(\cdot) > 0$ and $S'(\cdot) > 0$. If $\widehat{\Pi} < \frac{d(1)-1}{d(1)}S(1)$, there exists some $p' \in [\underline{p}, 1)$ such that $\widehat{\Pi} = \frac{d(p')-1}{d(p')}S(p')$. Then, a bank can *both* charge $R = \infty$, thus reaping $\Theta = \infty$, and attract all the entrepreneurs by giving them more than $\widehat{\Pi}$, as follows. Entrepreneurs' payoff from a deal (p, R) is $\Pi(p, R) = \frac{(d(p)-1)R}{d(p)R-1}S(p)$. It increases with p, decreases with R, and $\frac{d(p)-1}{d(p)}S(p) = \lim_{R\to\infty}\Pi(p, R)$. If the bank offers $p = p' + \epsilon < 1$ and $R = \infty$, the entrepreneurs coming to the bank get more than $\widehat{\Pi}$: $\Pi(\infty, p' + \epsilon) = \frac{d(p)-1}{d(p)}S(p)|_{p=p'+\epsilon} > \frac{d(p')-1}{d(p')}S(p') = \widehat{\Pi}$, where the inequality applies the fact that $\frac{d(\cdot)-1}{d(\cdot)}S(\cdot)$ is increasing as noted above.

Second, we show that $\Theta(K,\widehat{\Pi}) = \infty$ if $\widehat{\Pi} = \frac{d(1)-1}{d(1)}S(1)$. Let $f(\widehat{\Pi}, K; p) := (K\cdot(R(p;\widehat{\Pi})-1) - C(p)$. Then, $\Theta(K,\widehat{\Pi}) = \max_{p\in[0,1]} f(\widehat{\Pi}, K; p)$ and $f'_p = K\widehat{\Pi}\frac{(d(p)-1)S'(p)+d'(p)(S(p)-\widehat{\Pi})}{[d(p)\widehat{\Pi}-(d(p)-1)S(p)]^2} - C'(p)$. Of these two terms, at $p \approx 1$, the first one is in the order of $\frac{1}{(1-p)^2}$ if $\widehat{\Pi} = \frac{d(1)-1}{d(1)}S(1)$, whereas the second one, $C'(p)$, by assumption, is in the order of $o(\frac{1}{(1-p)^2})$, dominated by the first term. It follows that $f'_p > A\frac{1}{(1-p)^2}$ if $p > p_0$ for some $p_0 \in (0, 1)$ and $A > 0$. Then, $\lim_{p\to1} f(\widehat{\Pi}, K; p) = \lim_{p\to1}[f(\widehat{\Pi}, K; p_0) + \int_{p_0}^{p} f'_p dp] > f(\widehat{\Pi}, K; p_0) + A\cdot\lim_{p\to1}\int_{p_0}^{p}\frac{1}{(1-s)^2}ds = \infty$. That is, at $\widehat{\Pi} = \frac{d(1)-1}{d(1)}S(1)$, the value $\Theta(K,\widehat{\Pi}) = \infty$ and the optimal choice $\rho(K,\widehat{\Pi}) = 1$. ■

For Proposition 1:

Proof. (i) If $\Theta(K,\widehat{\Pi})>0$, then $K(R-1)>C(\rho(K,\widehat{\Pi}))>0$, therefore $R-1>0$. By (11) and the envelope theorem, then, $\partial\Theta(K,\widehat{\Pi})/\partial K = R(\rho(K,\widehat{\Pi}),\widehat{\Pi})-1>0$, and $\partial^2\Theta(K,\widehat{\Pi})/\partial K^2 = R'_p\cdot\partial\rho/\partial K$. We saw $R'_p>0$ if $\Theta>0$ at footnote 13. Therefore, it suffices to show $\partial\rho/\partial K>0$, for which denote the left-hand-side (LHS) term of (15) by $Y(p,\widehat{\Pi})$. Then, by the implicit function theorem, $\frac{\partial\rho(K,\widehat{\Pi})}{\partial K} = \frac{\partial Y}{\partial K}/[-(\frac{\partial Y}{\partial p}-C'')]$ and $\frac{\partial\rho(K,\widehat{\Pi})}{\partial\widehat{\Pi}} = \frac{\partial Y}{\partial\widehat{\Pi}}/[-(\frac{\partial Y}{\partial p}-C'')]$. The second order condition of the maximization problem (11) implies that $\frac{\partial Y}{\partial p}-C''<0$ at $p=\rho(K,\widehat{\Pi}):=\widetilde{p}$. Therefore, to prove $\partial\rho/\partial K>0$, it suffices to show $\frac{\partial Y}{\partial K}>0$, which holds true as Y is positively proportional to K.

(ii) By the envelope theorem $\partial\Theta(K,\widehat{\Pi})/\partial\widehat{\Pi} = K\frac{-(d(p)-1)S(p)}{[d(p)\widehat{\Pi}-(d(p)-1)S(p)]^2}|_{p=\widetilde{p}}<0$. To show $\partial\rho/\partial\widehat{\Pi}<0$, by the argument above, it suffices to prove that $\frac{\partial Y}{\partial\widehat{\Pi}}<0$. For this purpose, note that $\frac{Y}{K} = \frac{(d(\widetilde{p})-1)S'(\widetilde{p})\widehat{\Pi}}{[d(\widetilde{p})\widehat{\Pi}-(d(\widetilde{p})-1)S(\widetilde{p})]^2}+d'(\widetilde{p})\cdot\frac{(S-\widehat{\Pi})\widehat{\Pi}}{[d\widehat{\Pi}-(d-1)S]^2}$; both terms are to be shown decreasing with $\widehat{\Pi}$. For the first, $(d(\widetilde{p})-1)S'(\widetilde{p})>0$ and $\frac{\widehat{\Pi}}{[d(\widetilde{p})\widehat{\Pi}-(d(\widetilde{p})-1)S(\widetilde{p})]^2}$ decreases with $\widehat{\Pi}$ for $\widehat{\Pi}>\frac{d(1)-1}{d(1)}S(1)$ (by Lemma 1) and thus bigger than $\frac{(d(\widetilde{p})-1)}{d(\widetilde{p})}S(\widetilde{p})$. For the second, $d'(\widetilde{p})>0$ and $\{\frac{(S-\widehat{\Pi})\widehat{\Pi}}{[d\widehat{\Pi}-(d-1)S]^2}\}'_{\widehat{\Pi}}<0 \Leftrightarrow (S-2\widehat{\Pi})[d\widehat{\Pi}-(d-1)S]<2d(S-\widehat{\Pi})\widehat{\Pi}$. Note that $\Theta>0$ only if $S>\widehat{\Pi}$, namely if the surplus, $S-\widehat{\Pi}$, that the bank gets from each project screened is positive; and also note that $d\widehat{\Pi}-(d-1)S>0$ by Lemma 1. Therefore, the last inequality of the chain above holds true if $S-2\widehat{\Pi}<0$. If $S-2\widehat{\Pi}>0$, the left-hand-side of that inequality is smaller than $(S-2\widehat{\Pi})d\widehat{\Pi}<d(S-\widehat{\Pi})\widehat{\Pi}<2d(S-\widehat{\Pi})\widehat{\Pi}$, the-right-hand side.

■

For Proposition 2

Proof. (i): From the proof of Proposition 1, we know that $\partial\Theta(K,\widehat{\Pi})/\partial K = R-1>0$ if $\Theta(K,\widehat{\Pi})>0$. Thus, $\Theta(K,\widehat{\Pi})$ is an increasing function of K over the range where $\Theta(K,\widehat{\Pi})>0$. And $\Theta(K,\widehat{\Pi})\to\infty$ if $K\to\infty$. Therefore, $\underline{K}(\widehat{\Pi})=\inf\{K|\Theta(K,\widehat{\Pi})>0\}$ is well defined. By this definition and the increasing of Θ with K, $\underline{K}$ is the largest root of $\Theta(K,\widehat{\Pi})=0$ and $\Theta(K,\widehat{\Pi})>0$ if and only if $K>\underline{K}$.

We then proceed to show $\underline{K} > 0$. For this purpose, note that $\Theta(K, \widehat{\Pi}) = 0$ and $\rho(K, \widehat{\Pi}) = 0$ at $K = 0$. And also for $\widehat{\Pi} > \frac{d(1)-1}{d(1)} S(1)$, $\Theta(K, \widehat{\Pi})$ and $\rho(K, \widehat{\Pi})$ are continuous with K. By the continuity there exists some $\varepsilon > 0$ such that $\widetilde{p} := \rho(K, \widehat{\Pi}) < \underline{p}$ if $K < \varepsilon$. Then, for these K, $S(\widetilde{p}) < 0$. It follows that $R(\widetilde{p}, \widehat{\Pi}) = \frac{\widehat{\Pi}}{d(p)\widehat{\Pi} - (d(p)-1)S(p)}|_{p=\widetilde{p}} < 1$. Therefore, for $K < \varepsilon$, $\partial\Theta(K, \widehat{\Pi})/\partial K = R - 1 < 0$. As $\Theta(0, \widehat{\Pi}) = 0$, we have $\Theta(K, \widehat{\Pi}) < 0$ if $0 < K < \varepsilon$; that is, a bank of size K makes loss if it chooses to invest in screening expertise and compete for entrepreneurs. Therefore, $\underline{K}(\widehat{\Pi}) > \varepsilon > 0$.

(ii) First, $\partial\Theta(K, \widehat{\Pi})/\partial K|_{K=\underline{K}} = R - 1 > 0$: for banks with $K = \underline{K}$, they get 0 value if becoming informed; and thus $(R-1)\underline{K} = C(\rho(\underline{K}, \widehat{\Pi})) > 0$. Second, by Proposition 1(ii), $\partial\Theta(K, \widehat{\Pi})/\partial\widehat{\Pi} < 0$. Then, by the implicit function theorem, $\underline{K}'(\widehat{\Pi}) = -\frac{\partial\Theta(K,\widehat{\Pi})/\partial\widehat{\Pi}}{\partial\Theta(K,\widehat{\Pi})/\partial K}|_{K=\underline{K}} > 0$. ■

For Proposition 3:

Proof. To prove the existence and uniqueness of equilibrium, it suffices to show that (16) and (17) have a unique solution of $(t, \widehat{\Pi})$. For this purpose, a key role is played by the following claim, which is proved after the proposition.

Claim A: The mass of entrepreneurs served by a bank of size K, $\beta(K, \widehat{\Pi})$, given by (14), satisfies: $\partial\beta(K, \widehat{\Pi})/\partial\widehat{\Pi} < 0$, $\partial\beta(K, \widehat{\Pi})/\partial K > 0$, and $\beta(K, \widehat{\Pi}) \to \infty$ if $\widehat{\Pi} \to \frac{d(1)-1}{d(1)} S(1)$.

To simplify notation, let $\beta_j(\widehat{\Pi}) := \beta(K_j, \widehat{\Pi})$. With the claim, $g(\widehat{\Pi}) := \int_0^1 \beta_j(\widehat{\Pi})$ decreases with $\widehat{\Pi}$ and goes to ∞ if $\widehat{\Pi} \to \frac{d(1)-1}{d(1)} S(1)$. And obviously $\lim_{\widehat{\Pi}\to\infty} g(\widehat{\Pi}) = 0$, namely, if entrepreneurs demand a too large payoff no banks are willing to serve them. Therefore, $g(\widehat{\Pi}) = 1$ has a unique solution, denoted by $\widehat{\Pi}_a$. Two cases may arise.

Case 1: $K_1 \geq \underline{K}(\widehat{\Pi}_a)$. Then, the unique solution to the simultaneous equations of (16) and (17) is $t = 1$ and $\widehat{\Pi} = \widehat{\Pi}_a$.

Case 2: $K_1 < \underline{K}(\widehat{\Pi}_a)$. That is, not all banks become informed, namely, $t < 1$. Thus (16) is reduced to $K_t = \underline{K}(\widehat{\Pi})$. To this equation and equation (17) we show there is a unique

solution. By Proposition 2(ii), $\underline{K}(\widehat{\Pi})$ is a strictly increasing function. Therefore, the inverse function of $\underline{K}(\widehat{\Pi})$ exists, denoted by $\Psi(K)$; thus, $\Psi(K)$ is the level of entrepreneurs' payoff at which banks of size K are indifferent between being informed and staying out. From (16) $\widehat{\Pi} = \Psi(K_t)$. Substitute it into (17), which then becomes:

$$f(t) := \int_0^t \beta_j(\Psi(K_t)) = 1.$$

We now prove $f(t) = 1$ has a unique solution $t < 1$ by noting or showing the following four points. First, $f(\cdot)$ is continuous. Second, $f(0) = 0$. Third, $f'(t) > 0$ because $f'(t) = \beta_t(\Psi(K_t)) + \int_0^t \beta_j'(\widehat{\Pi}) \cdot \Psi'(K_t) \cdot dK_t/dt > 0$, since $\beta_j'(\widehat{\Pi}) < 0$ by Claim A, $dK_t/dt \leq 0$ by assumption, and $\Psi'(K_t) > 0$ as $\Psi(K)$ is an inverse function of $\underline{K}(\widehat{\Pi})$ and $\underline{K}'(\widehat{\Pi}) > 0$ by Proposition 2(ii). And fourth, $f(1) > 1$. In this case, $K_1 < \underline{K}(\widehat{\Pi}_a)$, which is equivalent to $\Psi(K_1) < \widehat{\Pi}_a$. Then $f(1) = \int_0^1 \beta_j(\Psi(K_1))|_{\beta_j'(\widehat{\Pi})<0} > \int_0^1 \beta_j(\widehat{\Pi}_a)|_{\text{definition of } \widehat{\Pi}_a} = 1$.

(i): That is, if for any $m \in \Omega$, K_m is increased, where Ω is a positively measured subset of $[0, t]$, then $\widehat{\Pi}$ goes up and for any $j \notin \Omega$ and $j \leq t$, R_j goes down. To prove the former result, we apply *reductio ad absurdum*. If, on the contrary, $\widehat{\Pi}$ non-increases, then, $\underline{K}$ non-decreases by Proposition 2, therefore the number of informed banks is not reduced, namely, the threshold t is not decreased. For each of the informed banks, say bank j, by Claim A, its market share, $\beta_j(\widehat{\Pi})$, non-decreases. But the market share of all banks in set Ω *strictly* increases because their capacities are enlarged. Moreover, set Ω has a positive measure. It follows that $\int_0^t \beta_j(\widehat{\Pi})$ is *strictly* increased, thus above 1, which contradicts the market clearing condition, (17).

To prove that for any $j \notin \Omega$, p_j and R_j both fall, first note that as $p_j = \rho(K_j, \widehat{\Pi})$ and $\partial\rho/\partial\widehat{\Pi} < 0$, with $\widehat{\Pi}$ rising and K_j fixed, p_j falls. Second, R_j as a function of p_j and $\widehat{\Pi}$, given by (10), increases with p_j and decreases with $\widehat{\Pi}$. Then, with p_j going down and $\widehat{\Pi}$ up (as shown above), R_j decreases.

(ii): To prove the result, it suffices to show that if $\widehat{\beta}_j$ is given, then $d\widehat{p}_j/dK_j < 0$. Note

that $\widehat{p}_j$ satisfies the first order condition, (15), with K replaced by K_j, namely,

$$K_j\widehat{\Pi}\frac{(d(p)-1)S'(p)+d'(p)(S(p)-\widehat{\Pi})}{[d(p)\widehat{\Pi}-(d(p)-1)S(p)]^2}=C'(p). \tag{29}$$

To get $\widehat{\Pi}$ as a function of K and β, note that for any informed bank $\Pi(p,R)=\widehat{\Pi}$, which, with $\Pi(p,R)$ given by (7), becomes:

$$S(p)\frac{(d(p)-1)R}{d(p)R-1}=\widehat{\Pi}. \tag{30}$$

Solving R from (5) as a function of β, and substituting it into (30), we find $\widehat{\Pi}$ related to K_j and $\widehat{\beta}_j$ through

$$\widehat{\Pi}=\frac{d(p)-1}{d(p)}(S(p)+\frac{K_j}{\widehat{\beta}_j})|_{p=\widehat{p}_j}. \tag{31}$$

Substitute it for $\widehat{\Pi}$ into (29), and then find $\widehat{p}_j$ solely determined by K_j through

$$\frac{\widehat{\beta}_j}{d(p)}(\frac{\widehat{\beta}_jS(p)}{K_j}+1)[S'(p)+\frac{d'(p)}{d(p)}(\frac{S(p)}{d(p)-1}-\frac{K_j}{\widehat{\beta}_j})]=C'(p). \tag{32}$$

To prove that given $\widehat{\beta}_j$, $d\widehat{p}_j/dK_j<0$, denote the LHS of the above equation by $X(p,K_j)$. Then, by the implicit function theorem, $\frac{d\widehat{p}_j}{dK_j}=\frac{\partial X}{\partial K_j}/[-(\frac{\partial X}{\partial p}-C'')]$. Obviously $\frac{\partial X}{\partial K_j}<0$, because $S(p)>0$ and $d'(\cdot)>0$. Therefore, it suffices to prove $\frac{\partial X}{\partial p}-C''<0$. Let $g(p,K):=\frac{d(p)-1}{d(p)}(S(p)+\frac{K}{\widehat{\beta}_j})$. Then, $\widehat{\Pi}=g(\widehat{p}_j,K_j)$ by (31). Thus, $X(p,K)=Y(p,g(p,K))$, where $Y(p,\widehat{\Pi})$ was used to denote the LHS of equation (15) (namely, 29 here) in the proof of Proposition 1. Therefore, $\frac{\partial X}{\partial p}=\frac{\partial Y}{\partial p}+\frac{\partial Y}{\partial\widehat{\Pi}}\frac{\partial g}{\partial p}<\frac{\partial Y}{\partial p}$, because $\frac{\partial Y}{\partial\widehat{\Pi}}<0$, as was shown in the proof of Proposition 1(ii), and $\frac{\partial g}{\partial p}>0$. It follows that $\frac{\partial X}{\partial p}-C''<\frac{\partial Y}{\partial p}-C''<0$, which was shown in the proof of Proposition 1(i).

For each informed bank j, as $\widehat{p}_j$ decreases and also $\widehat{\Pi}$ increases (by result i), the interest rate it charges, $R(\widehat{p}_j,\widehat{\Pi})$, decreases, as was noted above. ■

For Claim A, which is used for proving Proposition 3:

Proof. By (5), $\beta(K,\widehat{\Pi})=\frac{K}{E(K,\widehat{\Pi})}$, where $E(K,\widehat{\Pi}):=n_g(p)I(p,R)$, with p and R as functions of $(K,\widehat{\Pi})$ given respectively by (12) and (13). To prove that $\partial\beta(K,\widehat{\Pi})/\partial\widehat{\Pi}<0$

and $\partial\beta(K,\widehat{\Pi})/\partial K > 0$, it suffices to prove that $E'_{\widehat{\Pi}} > 0$ and $E'_K < 0$. Substituting (10) for R in $I(p,R)$ which is given by (4), we find $E = \frac{d}{d-1}\widehat{\Pi} - S := f(p,\widehat{\Pi})$. Note that d and S are both functions of p only, while $p = \rho(K,\widehat{\Pi})$ by (12). Therefore, $E'_{\widehat{\Pi}} = \partial f/\partial p \cdot \partial\rho/\partial\widehat{\Pi} + \partial f/\partial\widehat{\Pi}$ and $E'_K = \partial f/\partial p \cdot \partial\rho/\partial K$. Straightforwardly $\partial f/\partial\widehat{\Pi} > 0$. And by Proposition 1(ii), $\partial\rho/\partial\widehat{\Pi} < 0$. Then, $E'_{\widehat{\Pi}} > 0$ follows $\partial f/\partial p < 0 \Leftrightarrow (\frac{d}{d-1})'_p\widehat{\Pi} - S'(p) < 0|_{S'>0,\widehat{\Pi}>0} \Leftarrow (\frac{d}{d-1})'_p < 0 \Leftrightarrow -\frac{d'(p)}{(d-1)^2} < 0$ which holds true because $d' > 0$ by (3). Moreover, as $\partial f/\partial p < 0$, and $\partial\rho/\partial K > 0$ by Proposition 1(i), $E'_K = \partial f/\partial p \cdot \partial\rho/\partial K < 0$.

For $\lim_{\widehat{\Pi}\to\frac{d(1)-1}{d(1)}S(1)}\beta(K,\widehat{\Pi}) = \infty$, it suffices to prove that $I(p,R) = \frac{V_g(p)}{Rd(p)-1}|_{p,R \text{ given by (12), (13)}} \to 0$ if $\widehat{\Pi} \to \frac{d(1)-1}{d(1)}S(1)$, which follows from $\lim_{\widehat{\Pi}\to\frac{d(1)-1}{d(1)}S(1)} R(\rho(K,\widehat{\Pi}),\widehat{\Pi}) = \infty$. For this, note first that by the proof of Lemma 1, if $\widehat{\Pi} \to \frac{d(1)-1}{d(1)}S(1)$, then $p = \rho(K,\widehat{\Pi}) \to 1$. The LHS of (30) thus goes to $S(1)\frac{(d(1)-1)R}{d(1)R-1}$, while the right-hand side, namely $\widehat{\Pi}$, goes to $\frac{d(1)-1}{d(1)}S(1)$, which implies $R \to \infty$. ■

For Proposition 4:

Proof. To compare p_j^* with the equilibrium quality, $\widehat{p}_j$, rewrite equation (32) that characterizes $\widehat{p}_j$. By (10), $\widehat{R}_j = \frac{\widehat{\Pi}}{d(\widehat{p}_j)\widehat{\Pi}-(d(\widehat{p}_j)-1)S(\widehat{p}_j)}$. Substitute for $\widehat{\Pi}$ from (31), then,

$$\widehat{R}_j = \frac{1}{d(\widehat{p}_j)}(\frac{\widehat{\beta}_j S(\widehat{p}_j)}{K_j} + 1). \tag{33}$$

It follows that $S(\widehat{p}_j) = (\widehat{d}_j\widehat{R}_j - 1)K_j/\widehat{\beta}_j$. Substitute it for $S(p)$ in (32), and $\widehat{p}_j$ is then characterized by:

$$\widehat{R}_j(\widehat{\beta}_j S'(p) + \frac{d'(p)K_j}{d(p)-1}(\widehat{R}_j - 1)) = C'(p). \tag{34}$$

Now we come to compare p_j^* with $\widehat{p}_j$. Let $\eta := \frac{d'(\widehat{p}_j)K_j}{d(\widehat{p}_j)-1} > 0$ be a constant and $p(x)$ be the function implicitly defined by $U(x,p) := x(\widehat{\beta}_j S'(p) + \eta(x-1)) - C'(p) = 0$ for $x \geq 1$. Then, $p(1) = p_j^*$ and $p(\widehat{R}_j) = \widehat{p}_j$. As informed banks charge $\widehat{R}_j > 1$, the proposition is equivalent to $p'(x) > 0$. By the implicit function theorem, $p'(x) = -U'_x/U'_p$. Obviously, $U'_x > 0$. Moreover, $U'_p = x\widehat{\beta}_j S'' - C'' \leq -C'' < 0$, as $S'' \leq 0$ (by 3). Thus, $p'(x) > 0$. ■

For Lemma 2:

Proof. If out of $K+D$ units of funds under its deployment, the bank invests M in bad projects and $K+D-M$ in good ones, what the bank gets in each state is as follows.

In state ϕ, no projects succeed and the bank gets nothing.

In state 1, high-type projects succeed, but low types fail. By the LLN, out of all the good projects, the fraction of high types is $\Pr(\widetilde{q}=\overline{q}|\widetilde{s}=g) := h_g$, while the fraction out of the bad projects is $\Pr(\widetilde{q}=\overline{q}|\widetilde{s}=b) := h_b$. Hence, in state 1, fraction h_g of the investment in good projects and h_b of that in bad ones succeed. Success delivers a return rate F in the former investment and F' in the latter investment. Therefore, the revenue of the bank in state 1 is $(K+D-M)h_gF + Mh_bF' := Q(M)$. And its liability duty is Df. The bank might default in this state. Its profit is then $\max\{Q(M)-Df, 0\} := \Theta_1(M)$.

In state 2, all the bank's projects succeed. The bank does not default in the state, or it gets 0 profit, certainly not the case. Hence, the bank's profit is $(K+D-M)F + MF' - Df := \Theta_2(M)$.

Altogether, the expected profit of the bank is $\Theta(M) = (\overline{q}-\underline{q})\Theta_1(M) + \underline{q}\Theta_2(M)$. We show this function has two local maximizers, $M=0$ and $M=K+D$. If M is small enough such that $\Theta_1(M) \geq 0$, then

$$\Theta(M) = (K+D-M)F \cdot [(\overline{q}-\underline{q})h_g + \underline{q}] + MF' \cdot [(\overline{q}-\underline{q})h_b + \underline{q}] - D \cdot \overline{q}f.$$

Note that $(\overline{q}-\underline{q})h_g + \underline{q} = \overline{q}\Pr(\widetilde{q}=\overline{q}|\widetilde{s}=g) + \underline{q}(\widetilde{q}=\underline{q}|\widetilde{s}=g) = q_g$ and, similarly, $(\overline{q}-\underline{q})h_b + \underline{q} = q_b$. Therefore, $\Theta(M) = (K+D)q_gF - M(q_gF - q_bF') - D\cdot\overline{q}f$. It decreases with M because $q_gF > q_bF'$ by (19). Thus the maximum occurs at $M=0$. If M is so big that $\Theta_1(M)=0$, then $\Theta(M) = \underline{q}((K+D)F + M(F'-F) - Df)$. It increases with M because $F' > F$ by (19). Thus, the maximum occurs for this case at $M = K+D$.

To prevent the bank from investing in the evaluated bad projects, it commands $\Theta(0) \geq \Theta(K+D)$, which, by substituting (19) for F' and noting $F = R/q_g$, gives rise to (20).

And $\Theta(0) > \Theta(K+D)$ if and only if $\Theta_1(M) > 0$. That is, if the bank is prevented from risk-shifting, it repays the debt in both states 1 and 2, thus with probability $\overline{q}$. ■

For Proposition 5:

Proof. By (22), L increases with d and R, both in turn increasing with p. Therefore, $L'(p) > 0$. And by (23), $p'_j(L_j) = \partial\rho/\partial K \cdot K_j > 0$, because $\partial\rho/\partial K > 0$ by Proposition 1(ii). ■

University of Essex

Submitted on 27/07/2012

References

[1] Akhavein, J.D., Berger, A. N. and Humphrey, D. B. (1997). "The Effects of Megamergers on Efficiency and Prices: Evidence from a Bank Profit Function", *Review of Industrial Organization*, vol. 12, 95-139.

[2] Besanko, D., Doraszelski, U., Kryukov, Y. and Satterthwaite M. (2011). "Learning-by-Doing, Organizational Forgetting, and Industry Dynamics", *Econometrica*, vol. 78, 453–508.

[3] Besanko, D., and Kanatas, G. (1993). "Credit Market Equilibrium with Bank Monitoring and Moral Hazard", *Review of Financial Studies*, vol. 6(1), 213-32.

[4] Berger, A., Demsetz, R. and Strahan, P. (1999). "The Consolidation of the Financial Services Industry: Causes, Consequences, and Implications for the Future," *Journal of Banking and Finance*, vol. 23, 123–94.

[5] Berger, A., Kashyap, A. and Scalise, J. (1995). "The Transformation of the U.S. Banking Industry: What a Long, Strange Trip It's Been," *Brookings Papers on Economic Activity*, vol. 2, 54–219.

[6] Best, R. and Zhang, H. (1993). "Alternative information sources and the information content of bank loans", *Journal of Finance*, vol. 48(4), 1507-1522.

[7] Billett, M. T., Flannery, M. J. and Garfinkel, J. A. (1995). "The Effect of Lender Identity on a Borrowing Firm's Equity Return," *Journal of Finance*, vol. 50, 699–718.

[8] Budd, C., Harris, C. and Vickers, J. (1993). "A Model of the Evolution of Duopoly: Does the Asymmetry Between Firms Tend to Increase or Decrease?", *Review of Economic Studies*, vol. 60, 543–573.

[9] Cabral, L. (2011). "Dynamic Price Competition with Network Effects," *Review of Economic Studies*, vol. 78, 83–111.

[10] Cantillo, M. (2004), "A Theory of Corporate Fund Structure and Investment", *Review of Financial Studies,* vol. 17(4), 1103-1128.

[11] Carter, R. and Manaster, S. (1990). "Initial Public Offerings and Underwriter Reputation", *Journal of Finance*, Vol. 45 (4), pp. 1045-67.

[12] Dell'Ariccia, M. and Marquez, R. (2006). "Lending Booms and Lending Standards," *Journal of Finance*, 61, vol. 2511-2546.

[13] Demsetz, R. S. and Strahan, P. E. (1997) "Diversification, Size, and Risk at Bank Holding Companies", *Journal of Money, Credit, and Banking,* vol. 29, 300-313.

[14] Diamond, D. (1984). "Financial Intermediation and Delegated Monitoring", *Review of Economic Studies*, vol. 51 (3) 393-414.

[15] Flaherty, M. (1980). "Industry Structure and Cost-Reducing Investment." *Econometrica*, vol. 48, 1187-1209.

[16] Gale, D. and Hellwig, M. (1985). "Incentive-Compatibility Debt Contracts: The One-Period Problem", *Review of Economic Studies*, vol. 52, 647-663.

[17] Gilber, R. and Newbery, D. (1982). "Preemptive Patenting and the Persistence of Monopoly Power", *American Economic Review*, vol. 72, 514–526.

[18] Gorton, G. and Winton, A. (2003). "Financial Intermediation," in (G. Constantinides, M. Harris, and R. Stulz, eds.), *The Handbook of the Economics of Finance: Corporate Finance*, vol. 1 pp. 431-552, Amsterdam, the Netherlands: Elsevier Science.

[19] Hao, L. (2003). "Bank Effects and the Determinants of Loan Yield Spreads," working paper.

[20] Haq, M. and Heaney, R. (2012). "Factors determining European bank risk," *Journal of International Financial Markets, Institutions and Money*, vol. 22, 696–718.

[21] Hauswald, R. and Marquez, R. (2006). "Competition and Strategic Information Acquisition in Credit Markets," *Review of Financial Studies*, vol. 19, 967-1000.

[22] Holmstrom, B., and Tirole, J. (1997). "Financial intermediation, loanable funds, and the real sector", *Quarterly Journal of Economics*, vol. 112 (3), 663-691.

[23] Hortlund, P. (2005). "The Long-Term Relationship between Capital and Earnings in Banking: Sweden 1870–2001," SSE/EFI Working Paper Series in Economics and Finance 611

[24] James, C. (1987). "Some Evidence on the Uniqueness of Bank Loans, *Journal of Financial Economics*, vol. 19, 217-235.

[25] Jensen, M. C. and Meckling, W. H. (1976). "Theory of the firm: managerial behavior, agency costs and ownership structure", *Journal of Fianacial Economics,* vol. 3, 305-360.

[26] Jones, K. and Critchfield, T. (2005). "Consolidation in the U.S. Banking Industry: Is the 'Long, Strange Trip' About to End?" *FDIC Banking Review*, vol. 17(4), 31-61.

[27] Krasa, S. and Villamil, A. P. (1992). "Monitoring the Monitor: An Incentive Structure for a financial intermediary", *Jounal Economic Theory,* vol. 57 (1), 197-221.

[28] Kreps, D. and Scheinkman, J. (1983). "Quantity precommitment and Bertrand competition yields Cournot outcomes," *Bell Journal of Economics*, vol. 14, 326-337.

[29] Lepetit, L., Nys, E., Rous, P. and Tarazi, A. (2008). "Bank income structure and risk: An empirical analysis of European banks," *Journal of Banking and Finance,* vol. 32, 1452-1467.

[30] Liang, J. N. and Rhoades, S. A. (1991). "Asset Diversification, Firm Risk, and Risk-Based fund Requirements in Banking", *Review of Industrial Organization*, vol. 6, 49-59.

[31] Lummer, S. and McConnell, J. (1989). "Further evidence on the bank lending process and the fund-market response6 to bank loan agreements", *Journal of Financial Economics*, vol. 25, 99-122.

[32] Mikkelson, W. H. and Partch, M. M. (1986). "Valuation Effects of Security Offerings and the Issuance Process", *Journal of Financial Economics*, vol. 15, 91-118.

[33] Miles, D., Yang, J. and Marcheggiano, G. (2012). "Optimal Bank Capital," *Economic Journal*, doi: 10.1111/j.1468-0297.2012.02521.x.

[34] Poon, W, Lee, J. and Gup, B. (2009). "Do Solicitations Matter in Bank Credit Ratings? Results from a Study of 72 Countries," *Journal of Money, Credit and Banking*, vol. 41, 285–314.

[35] Rajan, R. (1994). "Why Bank Credit Policies Fluctuate: A theory and Some Evidence," *Quarterly Journal of Economics*, vol. 109, 399–441.

[36] Ross, D. (2010), "'The "Dominant Bank Effect:' How High Lender Reputation Affects the Information Content and Terms of Bank Loans," *Review of Financial Studies,* vol. 23 (7), 2730-2756.

[37] Rucks, M. (2004), "Bank Competition and Credit Standards," *Review of Financial Studies*, vol. 17, 1073-1102.

[38] Saunders, A. and Wilson, B. (1999), "The Impact of Consolidation and Safety-Net Support on Canadian, US and UK Banks: 1893–1992." *Journal of Banking and Finance*, vol. 23, 537–571.

[39] Townsend, R. (1979), "Optimal Contracts and Competitive Markets with Costly State Auditing", *Journal of Economic Theory,* vol. 21, 265-293.

[40] Williamson, S. D. (1986a). "Costly Monitoring, Financial Intermediation, and Equilibrium Credit Rationing", *Journal of Monetary Economics*, vol. 18, 159-179.

[41] Williamson, S. D. (1986b). "Increasing returns to Scale in Financial Intermediation and the Nonneutrality of Government Policy," *Review of Economic Studies*, vol. 53, 863-875.

[42] Winton, A. (1995). "Delegated Monitoring and Bank Structure in a Finite Economy", *Journal of Financial Intermediation,* vol. 4, 158-187.