On the welfare cost of bank concentration*

Sofía Bauducco[a] and Alexandre Janiak[b]

[a]Central Bank of Chile
[b]University of Chile

October 13, 2015

Abstract

We build a model of bank concentration. Banks and entrepreneurs meet in a credit market characterized by search frictions and negotiate repayment rates *à la* Nash. Banks are large in the sense that they allocate credit to more than one entrepreneur through branches and there is bank heterogeneity in terms of their cost structure. Banks have incentives to overlend, generating a *scale inefficiency* and overconcentration of banks. We find that this friction also generates too much concentration on the goods market, lowering aggregate output and welfare. We calibrate the model with data on the distribution of branches across banks in the US and available estimates on X-efficiency in the banking sector to assess the quantitative importance of this effect. We find that aggregate output would increase by 2.4% had the scale inefficiency been absent, while loan rates would decrease by 1.2%.

Keywords: Bank concentration; Bargaining; Search; Scale inefficiency; X-efficiency.
JEL codes: E44; G21; G28.

*We are grateful to Rui Albuquerque, Fernando Álvarez, Gabriele Camera, Ramiro De Elejalde, Elton Dusha, Eugenio Giolito, Ricardo Lagos, Eric Martínez Telchi, Toshi Mukoyama, Nicolas Petrosky-Nadeau, Carlos Ponce, Paulo Santos Monteiro, Shouyong Shi, Etienne Wasmer and Steve Williamson, as well as participants at the Shanghai Macroeconomic Workshop, the ENSAI workshop on "Frontiers in macroeconomics and finance", the CEA workshop on "Market imperfections and the macroeconomy", the SECHI Annual Meeting 2015 and seminars at Universitat Pompeu Fabra, PUC-Chile, Central bank of Chile, Ilades, USACH and University of Chile. We acknowledge funding from the Anillo in Social Sciences and Humanities (project SOC 1402 on "Search models: implications for markets, social interactions and public policy"). Alexandre Janiak also thanks Fondecyt (project no 1151053) and the Millennium Institute for Research in Market Imperfections and Public Policy. All errors are our own.

1 Introduction

Since the last economic downturn, some policy-makers and scholars have questioned the welfare consequences of bank concentration.[1,2] Yet, the empirical literature suggests an ambiguous relation between bank concentration and economic performance.[3] On the one hand, concentration may raise the profitability of some banks to the detriment of others, with negative consequences for social welfare. On the other hand, some banks may produce at more efficient scales than others, justifying high concentration.[4]

Related to the issue of bank concentration is the one of *overbranching*.[5] Figure 1 displays the evolution of the number of commercial banks and the average number of offices per bank since the 30's.[6] This characteristic of the US economy is not a consequence of the recent mergers and acquisitions brought about by the crisis: one can see from this graph that, while restrictions on branching were present in the United States until the seventies, the number of banks and the average number of offices per bank followed a relatively flat evolution over time.[7] However, over the last fifty years, the number of banks has decreased by a factor of three, while the number of offices per

[1] If we consider two economies, we refer to situations of larger 'bank concentration' in one of them when there are fewer banks and banks are larger with respect to the other economy. In the empirical literature, bank size has been measured by using variables such as deposits or loans. Measures of concentration are, for example, the Herfindahl-Hirschman index or the n-firm concentration ratio.

[2] Examples in the policy debate are the Independent Commission on Banking in the UK or the discussion about the implementation of a maximum interest rate in Chile. In the US, the Dodd-Frank Wall Street Reform and Consumer Protection Act imposes limits on concentration in the financial sector.

[3] Economic performance has been measured in several ways. At the micro level, the literature has considered bank profitability, deposit rates or loan rates, pass-through of monetary interest rates. At the macro level, aggregate growth has been considered, or even credit availability to SMEs. See Berger, Demirgüç-Kunt, Levine, and Haubrich (2004) and Degryse, Kim, and Ongena (2009) for reviews of the empirical literature. A discussion about possible mechanisms affecting bank concentration and efficiency can also be found in Cao and Shi (2001).

[4] A third channel by which bank concentration may impact economic performance is through its effect on the riskiness of the financial sector. Again, the effect of bank concentration through this channel seems to be ambiguous. While some authors claim that larger banks are able to diversify risk more effectively and, consequently, enhance the systemic stability of the financial sector, others state that large banks are usually deemed *too big to fail* and this has pervasive effects on the riskiness of their portfolios and on the sector as a whole. See Beck, Demirgüç-Kunt, and Levine (2007), Schaeck, Cihak, and Wolfe (2009) and Martinez-Miera and Repullo (2010), among many others, as examples of this strand of literature.

[5] See e.g. Berger, Leusner, and Mingo (1997).

[6] The data is from the Historical Statistics on Banking from the FDIC and can be downloaded from https://www2.fdic.gov/hsob/index.asp. Unfortunately, the database does not provide figures for the period 1967-1983. We thus resorted to interpolation for this interval in Figure 1.

[7] Most states in the US either prohibited branching during the fist half of the 20th century altogether or limited branching until the 1970s. Moreover, the Douglas Amendment to the 1956 Bank Holding Company Act prohibited the acquisitions of banks outside the state. Deregulations occurred mostly in the 1970s. See e.g. Cetorelli and Strahan (2006) and De Elejalde (2012) for more details.

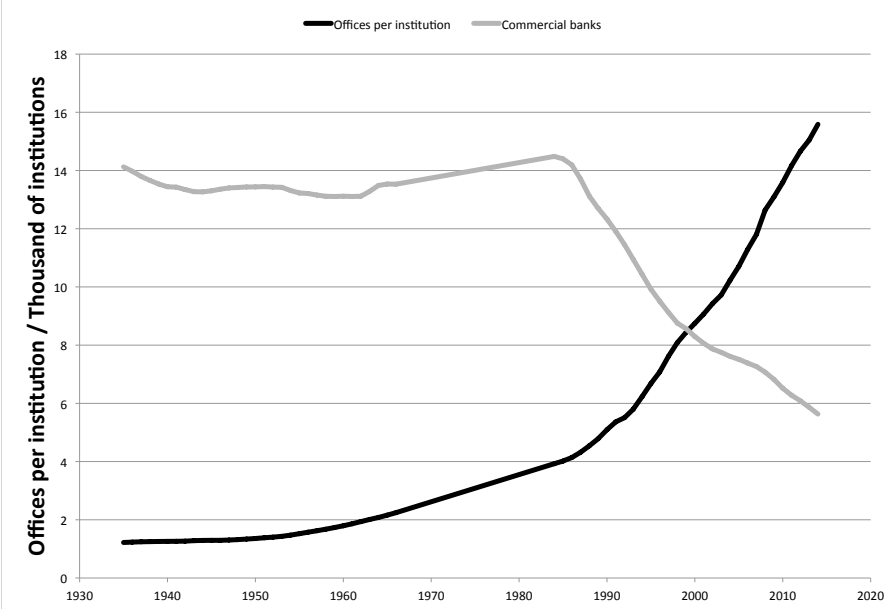

Figure 1: Number of commercial banks and offices per bank: 1937-2014

Source: Historical Statistics on Banking, FDIC.

commercial bank has increased from about 2 to 15. This change has been especially strong after the deregulations of the banking sector and has raised some concern in the policy debate (see Greenspan (2010)).

In this paper, we study the macroeconomic consequences of bank concentration in a search model of credit allocation. To allow for bank concentration, we depart from the credit search literature in two directions.[8] First, while it is common to assume that a bank allocates credit to one firm at most, we allow banks to have a continuum of customers, as banks can open more than one branch in our model. This allows us to study how bank size and the number of banks react to the economic environment. This is not an innocuous assumption, as it gives rise to strategic interactions between the bank and its customers in a context with bargaining. This has been discussed in Stole and Zwiebel (1996a) and Stole and Zwiebel (1996b).[9] In particular, banks can transfer part of the marginal cost of credit to the rate paid by firms through bargaining. Hence, in a context *à la* Lucas (1978), where the cost of credit is an increasing and convex function of the number of customers, banks have incentives to

[8] Wasmer and Weil (2004), den Haan, Ramey, and Watson (2003), and Diamond (1990), among others, have relied on search frictions to model the process of credit allocation.

[9] This is known as 'intra-firm bargaining' in the literature. In a context with search frictions, this inefficiency has been studied in Bertola and Caballero (1994), Smith (1999), Cahuc, Marque, and Wasmer (2008), among others.

overlend because they can negotiate higher repayment rates. Moreover, because banks operate at an inefficiently too large scale, some banks are forced out of the market. As a consequence, there is too much bank concentration in the sense that, compared to the efficient allocation, banks are larger and there are fewer banks.[10] We refer to this friction as *scale inefficiency*, an expression often used in the empirical literature on banking to characterize situations where banks are away from the minimum of their average cost curve.

Second, to account for the fact that some banks may be larger because they are more efficient, we assume that banks are heterogeneous in this dimension. We borrow from the macro literature on firm dynamics to model the distribution of efficiencies across banks (e.g. Hopenhayn (1992) and Melitz (2003)). Banks draw an efficiency parameter when entering the market and then decide to stay or leave. If they choose to stay, they operate at this efficiency level indefinitely. This generates an endogenous distribution of bank sizes that is influenced by the dispersion of efficiency draws and the convexity of the credit cost function.

We calibrate the model using data on the distribution of branches across banks in the United States, as well as available estimates of *X-efficiency* in the banking sector. This latter measure identifies the dispersion of efficiencies across banks in the sector. We then run some counterfactual experiments. Specifically, we study how aggregate performance is affected by the scale inefficiency present in the model. We find that, on top of generating too much bank concentration, the overlending behavior of banks entails too much concentration on the goods market as well. The mechanism behind this result is as follows: since bank concentration increases the cost of credit, it also reduces the number of firms in the economy, as potential entrepreneurs will be less willing to enter the market. Entrepreneurs reallocate to the labor market, generating an increase in labor supply, which depresses the equilibrium wage. As a consequence of the lower wage, incumbent firms choose to increase their size, thus increasing concentration on the goods market. Since the production function in the goods sector displays decreasing returns to scale, a more concentrated goods market implies lower aggregate output. Our quantitative exercise suggests that aggregate output would be 2.4% larger, had the scale inefficiency been absent, while loan rates would be 1.2% lower, resulting in an increase in aggregate welfare by 4.7%.[11]

Our mechanism is consistent with the "overbranching" behavior by banks in the

[10] There are two other sources of inefficiency in the model: congestion externalities *à la* Hosios (1990) that are not internalized in the negotiation process between a bank and an entrepreneur, and a hold-up problem on the credit market as in Acemoglu and Shimer (1999).

[11] We also show that most of the inefficiencies from the calibrated economy are due to the scale inefficiency, while congestion externalities and the hold-up problem seem to be minor.

United States documented by Berger, Leusner, and Mingo (1997), who showed that half of the branches of commercial banks were not profitable at the time of their analysis. According to these authors, *"banks prefer to open extra branches and operate on the upward-sloping portion of their average cost curve, experiencing scale diseconomies, because they receive extra revenues that offset the extra costs"*. Moreover, the general-equilibrium impact of overbranching on the goods market is consistent with the strand of literature showing that financial development eases competition and the entry of small firms, including Midrigan and Xu (2014), Guiso, Sapienza, and Zingales (2004), Cetorelli and Strahan (2006), Beck, Demirgüç-Kunt, Laeven, and Levine (2008), among others.

A key assumption in our model is continuous renegotiation of the repayment rate. Because of renegotiation, borrowers can potentially appropriate part of the financial contribution of lenders through a standard hold-up problem. Moreover, lenders can act strategically and decide to overbranch in order to renegotiate higher rates in case renegotiation occurs. While renegotiation is a standard assumption in the search literature, it is actually a fairly realistic assumption in the context of credit markets. Indeed, Roberts and Sufi (2009) document that over 90% of long-term debt contracts between firms and financial institutions are renegotiated prior to their stated maturity. This figure increases to 96% for contracts with stated maturity in excess of three years. Roberts and Sufi (2009) also show that renegotiation occurs relatively early and is typically not related to default or financial distress.

A body of literature, reviewed in Nosal and Rocheteau (2011), studies credit markets characterized by search frictions and buyers who have some private information about their ability to repay their debt. These papers include Aiyagari and Williamson (1999), Corbae and Ritter (2004), Kocherlakota (1998), Kocherlakota and Wallace (1998) and Shi (2005), among others. They discuss when efficiency can be restored in spite of adverse selection and lack of commitment. In Gu, Mattesini, Monnet, and Wright (2013), the lack of committment over debt repayment justifies the existence of banks as financial intermediaries. We think of our paper as complementary to this strand of literature. While we do not model in detail the behavior of borrowers, we focus on the inefficiencies generated by the behavior of lenders.

Corbae and D'Erasmo (2013) study the entry and exit decisions of banks along the business cycle, and how these decisions impact riskiness of the financial sector, and ultimately the real economy. To this end, they set up a model of banking industry dynamics in which an endogenous distribution of banks arises, and embed this market structure within a general equilibrium macroeconomic model. In this respect, our approach and theirs are related. The main difference between the two papers is that we do not address the issue of how the market structure in the banking sector affects

risk taking. We are also silent about how the market structure varies with the economic cycle. Instead, we focus our analysis on the interaction between bank concentration and concentration in the goods market, and the consequences of this interaction on aggregate macroeconomic variables such as output. Moreover, we discipline our model using data on the distribution of branches and the available estimates of efficiency in the banking sector, whereas Corbae and D'Erasmo (2013) mainly focus on loan data.

The rest of the paper is organized as follows. We present the model and characterize its equilibrium in Section 2. Comparative statics on the scale inefficiency are derived in Section 3. Section 4 describes the constrained-efficient allocations and compares them with the allocations of the decentralized equilibrium. The calibration strategy is discussed in Section 5. This section also describes the quantitative exercises performed and the main quantitative results of the paper. Section 6 concludes.

2 The model

2.1 Firms and workers

Time t is continuous and discounted at a rate $r > 0$. We study the steady state of an economy composed by a unit mass of agents, who can choose to be entrepreneurs or workers. Workers supply one unit of labor in the competitive labor market, where they earn a wage w. Entrepreneurs transit through two states. They first need to raise funds to create a firm. Such entrepreneurs are thus called *fund raisers*. Once the firm is created, they become *active entrepreneurs*; production starts and they earn profits.

The lifetime utility of an agent that faces the decision of being a worker or a fund raiser is denoted by V:

$$V = \max\{W, E\},$$

where E is the value for a fund raiser, and W is the value for a worker. In equilibrium, an agent is indifferent between these two options. Hence, the following no-arbitrage condition holds:

$$W = E. \tag{1}$$

Therefore, W can be written as

$$rW = w. \tag{2}$$

Fund raisers search for funds on the credit market. This market is characterized by search and matching frictions. Loans are provided by banks through branches, which are also active searchers on the market. Each branch finances at most one firm and a bank may own several branches. Denote by $\mathcal{E}$ the mass of fund raisers on the

market and $\mathcal{K}$ the total mass of branches searching for a partner. A matching function $m(\mathcal{E}, \mathcal{K})$ determines the mass of firms being actually financed at each point in time. This function follows standard properties: it is increasing in both arguments, concave, displays constant returns to scale and follows the property $m(\mathcal{E}, 0) = m(0, \mathcal{K}) = 0$.

The rate at which fund raisers find funds is p, and it can be computed as the share of matched entrepreneurs among all fund raisers:

$$p(\phi) = \frac{m(\mathcal{E}, \mathcal{K})}{\mathcal{E}} = m(1, \phi^{-1})$$

Because m has constant returns to scale, the rate p negatively depends on $\phi \equiv \frac{\mathcal{E}}{\mathcal{K}}$, which is called the *credit-market tightness*.

Denote by Π the value for an active entrepreneur. We can write E as

$$rE = p(\phi)[\Pi - E]. \tag{3}$$

For a firm to be created, it is required that a start-up cost κ is paid, and production starts immediately starts after. When an entrepreneur is assigned a loan, the bank with which the entrepreneur has been matched finances the start-up cost. In return, each time production takes place, the entrepreneur pays back an amount ρ to the bank. An active firm is destroyed at an exogenous rate $\lambda > 0$. When this occurs, the entrepreneur defaults on the loan. He then comes back to the pool of fund raisers and decides whether to search again for new funds or to become a worker.

Labor is the only factor required in the production process. All firms produce the same homogeneous final good, and we denote by $g(n)$ the production function of this good. It is strictly increasing, strictly concave and satisfies standard Inada conditions. Firms sell the produced good on a competitive market, the price of which is normalized to one.[12] On the cost side, entrepreneurs pay wage bills and the repayment flow ρ. Hence, the value for an active entrepreneur is

$$r\Pi = \max\{rV, \max_n g(n) - wn - \rho + \lambda[V - \Pi]\}. \tag{4}$$

Denote by n^* the optimal size chosen by active entrepreneurs. Because the labor

[12] An alternative is to consider that firms have monopolistic power over the variety they produce in a Dixit-Stiglitz fashion together with a linear technology. In this economy, the comparative statics of a policy change would be similar to the case with perfect competition and decreasing returns to scale, as long as the revenue function of firms in the monopolistic competition economy has the same curvature as the production function of the perfect competition economy. See Janiak and Santos Monteiro (2011) for an example illustrating this equivalence.

market is competitive, firm size is independent of the level of repayment flow ρ:

$$g'(n^*) = w. \tag{5}$$

By combining equations (2) to (5), we obtain the following increasing relation between firm size and credit-market tightness (for a given ρ):

$$\frac{g'(n^*)}{p(\phi)} = \frac{\pi(n^*) - \rho}{r + \lambda}, \tag{6}$$

where

$$\pi(n) = g(n) - g'(n)(n+1) \tag{7}$$

with $\pi'(n) = -(n+1)g''(n) > 0$.[13]

The left-hand side of (6) is the opportunity cost that a fund raiser faces to create a firm: it is equal to the wage the fund raiser would earn on the labor market multiplied by the time it takes on average to find funds (this latter is simply the inverse of the rate p). The right-hand side of (6) refers to the discounted sum of profits the fund raiser will earn once he becomes an active entrepreneur.[14]

2.2 Banks

At each point in time, banks open a mass of branches searching for projects to finance. Each branch that is not matched to an entrepreneur implies a flow opportunity cost $\eta > 0$. Matching occurs at a rate q:

$$q(\phi) = \frac{m(\mathcal{E}, \mathcal{K})}{\mathcal{K}} = \phi p(\phi),$$

an increasing function of ϕ.

Denote by M the stock of active firms from which a given bank receives payments. Managing those customers generates agency costs *à la* Lucas (1978) for the bank. As a consequence, the bank's profits are reduced by an amount $C_\varphi(M) = \frac{C(M)}{\varphi}$ at each point in time.[15] $C(M)$ is an increasing convex function of M, homogenous of degree $\alpha > 1$, that satisfies the property $C(0) = 0$ and the Inada conditions. The parameter φ refers to the efficiency of the bank, which is an exogenous characteristic. Banks are

[13] The fact that equilibrium profits are increasing in equilibrium firm size is due to the underlying variation in the equilibrium wage: when the wage decreases, profits increase and at the same time firm size increases.

[14] Notice that, on top of subtracting the labor cost from revenues, the formulation of profits (7) also takes into account the opportunity cost of an entrepreneur, which is equal to the wage rate. This explains the presence of the $(n+1)$ term in equation (7).

[15] Microfoundations for such a cost function can be found in Sealy and Lindley (1977).

heterogenous in this regard, and a higher value for φ describes a more efficient bank. Section 2.5 below describes in detail how the distribution of efficiencies is identified. In addition to agency costs, banks also face a fixed operating cost c.[16]

Finally, the bank has to finance the start-up cost of the recently matched projects. We denote by K the mass of branches searching for a partner. Hence, $Kq(\phi)$ branches are matched at each point in time. The respective flow cost is equal to $\kappa Kq(\phi)$.

Denote by $B(M;\varphi)$ the sum of discounted profits of a bank with efficiency φ and M active customers, which can be written as follows:

$$B(M;\varphi) = \max_{K} \frac{1}{1+rdt}\left[(\rho(M)M - \eta K - C_{\varphi}(M) - c)\,dt + B(M';\varphi) - \frac{\kappa K \phi p(\phi) dt}{1+rdt}\right] \tag{8}$$

such that

$$\dot{M} = K\phi p(\phi) - \lambda M, \tag{9}$$

where $dt \to 0$ is the size of an arbitrarily small interval of time. We use the prime notation to distinguish variables evaluated at time $t+dt$ from variables evaluated at time t.[17]

2.3 Bargaining

Once a branch and a fund raiser meet, they bargain over the repayment flow ρ under the following Nash rule:

$$\rho = \arg\max_{\rho} [\Pi - E]^{1-\beta}\left[\frac{\partial B}{\partial M} - \kappa\right]^{\beta} \tag{10}$$

The parameter $\beta \in (0,1)$ denotes the bargaining power of the bank, while $[\Pi - E]$ is the surplus of the entrepreneur and $\left[\frac{\partial B}{\partial M} - \kappa\right]$ the surplus of the bank. Production can start only if they agree upon a value for ρ. Renegotiation is allowed continuously, but, given that the κ cost is sunk, banks suffer from a hold-up problem. The following rule thus applies in case of renegotiation:

[16] In the literature on banking, economies of scale may result from several phenomena: regulation in the banking sector, liquidity insurance as in Diamond and Dybvig (1983), adverse selection and signaling as in Leland and Pyle (1977) or monitoring as in Diamond (1984), among others.

[17] The model could be extended by allowing banks to die at an exogenous rate λ_b. In this case, the allocations in the decentralized equilibrium would be the same as in the model without bank death if we introduced a market for annuities as in Blanchard (1985). A firm would buy the portfolio of customers of all dying banks and sell it to the surviving ones. This would provide insurance for banks, rendering their behavior invariant to λ_b.

$$\rho = \arg\max_{\rho} \left[\Pi - E\right]^{1-\beta} \left[\frac{\partial B}{\partial M} - \theta\kappa\right]^{\beta}. \tag{11}$$

The rule (11) differs from (10) through the presence of the parameter $\theta \in [0, 1]$, which reflects the stringency of the hold-up problem. For a value of $\theta = 1$, this friction is absent, while the hold-up problem is extreme when $\theta = 0$.[18]

If the parties renegotiate, production can continue only if they agree upon a new value for ρ. Production stops if the λ shock occurs, in which case the relation disappears as well as any specificity involved.

The rent in (11) for an active entrepreneur is

$$\Pi - E = \frac{\pi(n^*) - \rho}{r + \lambda}, \tag{12}$$

the discounted sum of profits, while, because renegotiation is possible, the surplus for the bank reads as

$$\frac{\partial B}{\partial M} = \frac{\rho(M) + \rho'(M)M - C'_{\varphi}(M)}{r + \lambda}. \tag{13}$$

The repayment flow depends on the mass of partners M the bank is involved with. Hence, the surplus for the bank is equal to the discounted sum of repayment flows net of the marginal agency cost, adding the effect of M on the renegotiated rates with all partners.

Similarly, the first-order condition of program (8) is:

$$\kappa + \frac{\eta}{\phi p(\phi)} = \frac{\rho + \rho'(M)M - C'_{\varphi}(M)}{r + \lambda}, \tag{14}$$

where the left-hand side of this expression is the cost of matching a branch to an entrepreneur, and the right-hand side is the surplus (13). The cost includes two components: the start-up cost financed by the bank and the search opportunity cost.

Notice that, in the first-order condition (14), the bank strategically chooses the mass of branches it opens, as this allows it to negotiate a higher ρ with all its partners.

The solution to (11) reads as

$$\rho = \beta\pi(n^*) + (1 - \beta)C'_{\varphi}(M) + (1 - \beta)(r + \lambda)\theta\kappa - (1 - \beta)\rho'(M)M. \tag{15}$$

[18] Two examples (among others) of microfoundations for θ are the following. First, it is common that, when firms default on a loan, banks may recover a fraction of the capital lent. The θ parameter can be interpreted as the recovered fraction in this case. Second, we could consider punishment for entrepreneurs who default without a justifiable cause. In this case, as long as the punishment is increasing in the loan κ, one can rewrite the resulting rule for ρ as in (11). See Kehoe and Levine (1993) and Banerjee and Newman (1993).

To understand this equation, notice that the entrepreneur would not accept a value for ρ higher than his profits, $\pi(n^*)$. The bank ideally would consider a value higher than the cost $(C'_\varphi(M) + (r+\lambda)\kappa)$ the relationship implies, but, given the nature of the hold-up problem, it can only claim a share θ of the setup cost κ. These are thus limits between which ρ has to be and the bargaining power of the bank β defines how close to $\pi(n^*)$ and how far from the bank's threat point the value of ρ is. This explains the presence of the first three terms in (15). The last term in (15) is due to the fact that the entrepreneur knows that, by being involved in a credit contract with the bank, the latter can renegotiate the repayment flow with its other customers. Then, the entrepreneur appropriates a share $(1-\beta)$ (the bargaining power of the entrepreneur) of the increase in the other customers' flow.

The solution to (10) is similar as it simply considers a value of $\theta = 1$:

$$\rho = \beta\pi(n^*) + (1-\beta)C'_\varphi(M) - (1-\beta)\rho'(M)M + (1-\beta)(r+\lambda)\kappa. \tag{16}$$

Because there is constant renegotiation and time is continuous, the value for ρ in (16) is not relevant for the equilibrium conditions as it is paid during an infinitely small amount of time.

2.4 Scale inefficiency

Equation (15) describes a differential equation in ρ. The following proposition provides the solution to this differential equation, as well as a version of the first-order condition (14) where the solution for ρ is integrated:

Proposition 1. *The equilibrium repayment flow ρ can be rewritten as*

$$\rho = (1-\beta)\varsigma + \beta\pi(n^*) + (1-\beta)(r+\lambda)\theta\kappa \tag{17}$$

with

$$\varsigma \equiv \Delta C'_\varphi(M) \tag{18}$$

and

$$\Delta = \frac{1}{\beta + \alpha(1-\beta)} \in (0,1), \tag{19}$$

implying the following first-order condition for the bank:

$$\left[1 - (1-\beta)\theta\right]\kappa + \frac{\eta}{\phi p(\phi)} = \beta\frac{\pi(n^*) - \varsigma}{r+\lambda}. \tag{CC}$$

Proof. See Appendices A.1 and A.2. □

The variable Δ is an *overlending factor*, generating a scale inefficiency in bank lending.[19] Indeed, if we compare in partial equilibrium (i.e., for given ϕ and n^*) the equilibrium allocation (CC) with the equilibrium allocation in an economy where a bank takes ρ as given when deciding on K, the equilibrium value for M is higher in the first economy: it can be shown easily that the first-order condition in that alternative economy would be the same as in (CC) with $\Delta = 1$. The intuition for overlending is the following. The repayment flow ρ is an increasing function of M because part of the marginal agency cost $C'_\varphi(M)$—an increasing function of M too—is passed on to the repayment flow charged to all loans through Nash bargaining. The bank knows it can influence the outcome of the bargaining process by varying *ex ante* the number of partners M. Hence, it chooses to assign an excessive amount of credit in order to obtain a higher value for ρ.

Notice that the lower the value for Δ, the more a bank overlends. Δ is a decreasing function of the curvature α, as the more convex the agency cost function $C(M)$ is, the more sensitive with respect to M the repayment flow ρ is, increasing the incentives for the bank to overlend. Similarly, Δ is increasing in the bank's bargaining power β. To understand this result, remember the discussion about equation (15): the higher the bargaining power of the entrepreneur, the more ρ depends on $C'_\varphi(M)$. The bank thus has more incentives to overlend when the entrepreneur's bargaining power is large, in order to influence ρ.

Because entrepreneurs can appropriate part of the change in the other customers' repayment flow, the variable Δ also appears in equation (17). However, we show in Section 3 that entrepreneurs end up paying higher repayment flows than what they would pay in an economy where agents consider ρ as given (i.e. in that economy Δ takes value one). The following corollary shows that, in spite of bank heterogeneity, all entrepreneurs pay the same ρ:

Corollary 1. *ς and ρ are independent of φ. Hence, all banks share the same ς and the same ρ.*

The proof of this corollary is straightforward. It is based on equation (CC) and the following version of the no arbitrage condition (6) for the entrepreneur where the equilibrium value for ρ given by equation (17) is integrated:

$$\frac{g'(n^*)}{p(\phi)} = (1-\beta)\left[\frac{\pi(n^*)-\varsigma}{r+\lambda} - \theta\kappa\right]. \tag{FC}$$

Indeed, given that all banks face the same tightness ϕ and the same firm size n^*, they must all be characterized by the same ς, implying the same ρ.

[19] See Section 4 for a more precise analysis of welfare.

We interpret ς as a measure of the inefficiency in the credit market. From Proposition 1, it is immediate to see that it is influenced by the size of the scale inefficiency. To determine the equilibrium value of ς, one also needs to determine the distribution of all M across all banks. We explore this issue in the following section.

2.5 Bank heterogeneity

In Sections 2.2 to 2.4, we have considered the optimal behavior of a given incumbent in the banking sector. In this section, we extend the analysis to the study of bank entry and exit. This allows us to identify the distribution of banks in the economy.

There is an infinite mass of potential entrants in the banking sector. Entry requires the payment of a sunk cost ν. Except for this cost, entry is free. Once the cost is paid, the efficiency parameter φ characterizing the new entrant is revealed: it is drawn from a distribution with continuous cumulative distribution function F. The density $f(\varphi) = F'(\varphi)$ has positive support over $(0, \infty)$. The bank can choose to exit if, given its draw of φ, it earns negative profits. We denote by φ^* the efficiency below which a bank chooses to exit. If the bank stays, it can start opening its first branches and operate as described in Sections 2.2 to 2.4.

Consider the economy in steady state. We show in Appendix A.3 that, once a bank enters and chooses to stay, it opens a sufficiently large mass of branches such that its mass of customers M immediately jumps to its long-run value. This immediate adjustment is the result of the linear structure of the search costs in the model. This is a convenient property because it allows us to calculate easily the value of an entering bank.

Denote by $M(\varphi)$ the long-run mass of customers of a bank with efficiency φ. The value of an entering bank that has just paid the sunk cost can be written as

$$rB(0;\varphi) = R(\varphi) - c, \tag{20}$$

where

$$R(\varphi) \equiv \Delta C'_{\varphi}(M(\varphi))M(\varphi) - C_{\varphi}(M(\varphi)) \tag{21}$$

is a measure of bank flow profits before the fixed operating cost c is inputed. Equation (20) is derived in Appendix A.3.

The following lemma shows that, although banks share the same ρ, more efficient banks are larger, i.e. they have more customers and earn larger profits:

Lemma 1. *Consider two firms with productivity φ_1 and φ_2, respectively. Then,*

$$M(\varphi_2) = M(\varphi_1)\left(\frac{\varphi_2}{\varphi_1}\right)^{\frac{1}{\alpha-1}} \tag{22}$$

and

$$R(\varphi_2) = R(\varphi_1)\left(\frac{\varphi_2}{\varphi_1}\right)^{\frac{1}{\alpha-1}}. \tag{23}$$

Proof. See Appendix A.4. □

To identify the equilibrium distribution of banks, we follow Melitz (2003) and call

$$\tilde{\varphi} = \left[\int_{\varphi^*}^{\infty} \varphi^{\frac{1}{\alpha-1}} \frac{dF(\varphi)}{1-F(\varphi^*)},\right]^{\alpha-1} \tag{24}$$

a measure of average efficiency in the banking sector. This variable is increasing in the efficiency threshold φ^* above which banks choose to stay in the sector. Given this definition, the following proposition identifies the equilibrium threshold φ^* and the average bank value:

Proposition 2. *The free-entry condition for banks can be written as*

$$\nu = [1 - F(\varphi^*)]\,\tilde{B}, \tag{FE}$$

while the zero-cutoff profit condition follows

$$\tilde{B} = \frac{c}{r}\left[\left(\frac{\tilde{\varphi}}{\varphi^*}\right)^{\frac{1}{\alpha-1}} - 1\right], \tag{ZCP}$$

where $\tilde{B}$ is the average bank value of an entrant:

$$\tilde{B} = \int_{\varphi^*}^{\infty} B(0;\varphi)\frac{dF(\varphi)}{1-F(\varphi^*)} \tag{25}$$

and also the value of a bank with efficiency parameter $\tilde{\varphi}$:

$$\tilde{B} = B(0;\tilde{\varphi}). \tag{26}$$

Proof. See Appendices A.5 and A.6. □

2.6 Aggregate inefficiency in the banking sector

Two elements of the model influence ς, the measure of inefficiency in the banking sector. The scale inefficiency described in Section 2.4 is a first component. Indeed, the overlending behavior by banks inflates the agency cost of handling more customers, decreasing the efficiency of the banking sector. A second component is bank selection, as described in Section 2.5. Depending on the cost structure of banks, the average efficiency of banks that survive may vary. This element influences the value of the threshold φ^*, as well as the average bank efficiency $\tilde{\varphi}$, as shown by equation (24).

The following proposition describes formally the discussion above as it shows that the equilibrium value of ς depends on both Δ and φ^*:

Proposition 3. *The measure of inefficiency in the banking sector ς can be written as*

$$\varsigma = \frac{\Delta C'(1) c^{\frac{\alpha-1}{\alpha}}}{[\Delta C'(1) - C(1)]^{\frac{\alpha-1}{\alpha}}} \varphi^{*-\frac{1}{\alpha}}. \tag{27}$$

Proof. See Appendix A.7. □

From Proposition 3, it is easy to see that ς is decreasing in both Δ and φ^*, that is, the banking sector becomes more efficient as Δ and φ^* increase.

2.7 Equilibrium

The free-entry condition (FE) and the exit condition (ZCP) allow us to identify the average bank value $\tilde{B}$ and the efficiency threshold φ^*. Figure 2 displays the two curves. The free-entry condition is increasing in the space $(\varphi^*, \tilde{B})$, while the zero-cutoff profit condition is decreasing. Intuitively, a high value for φ^* suggests low survivability for banks. Hence, in equilibrium, expected bank profits have to be large for new banks to be willing to enter, explaining the positive slope of the (FE) relation. The negative slope of the (ZCP) relation is the result of cost pressure for banks. Indeed, when costs are high, it is natural that only the most productive banks survive and that profits are low. This explains the negative relation between φ^* and $\tilde{B}$ along the (ZCP) locus.

Melitz (2003) shows that the (ZCP) curve cuts the (FE) once from above, implying that an equilibrium value for φ^* exists and is unique. Given the value for φ^*, the equilibrium value of $\tilde{\varphi}$ follows directly from equation (24). The value for Δ is also obtained directlty from equation (19). Given these values, we obtain the measure of inefficiency of the banking sector ς from equation (27).

With the equilibrium value for ς in hand, conditions (CC) and (FC) allow us to identify ϕ and n^*. Figure 3 displays the two loci. The firm creation condition is

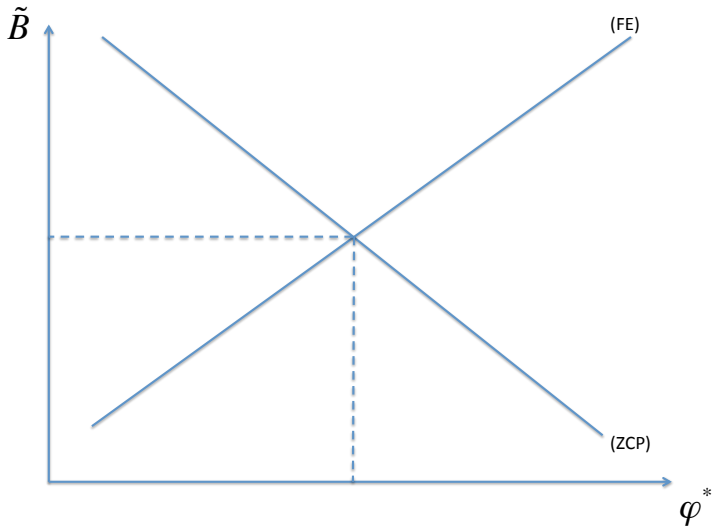

Figure 2: Identification of the average bank value $\tilde{B}$ and efficiency threshold φ^* through the free-entry and zero-cutoff profit conditions.

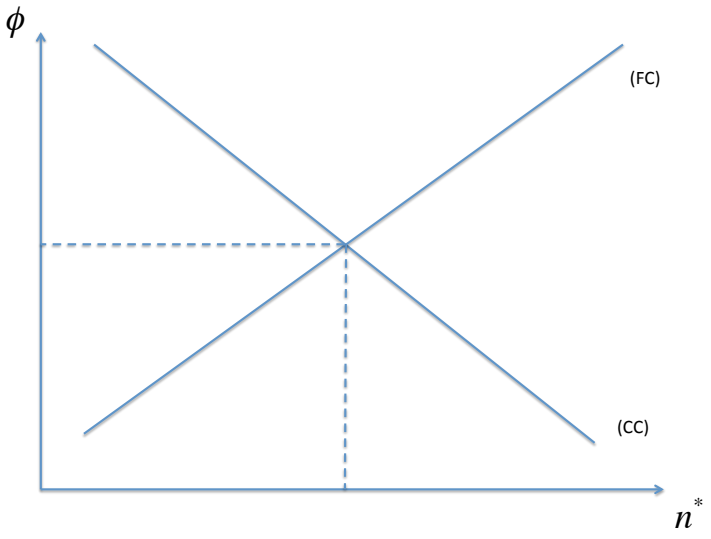

Figure 3: Identification of the tightness ϕ and firm size n^* through the firm creation and credit creation conditions.

increasing in the (n^*, ϕ) space. Intuitively, when n^* is large, the wage an entrepreneur could get on the labor market is low, providing incentives to entrepreneurs to start creating new firms (implying a large value for ϕ). At the same time, when n^* is large, expected profits resulting from firm creation are large, which reinforces the previous effect. On the other hand, the slope of the credit creation condition is negative. The intuition for this is as follows: because firm profits are high when n^* is large, banks' profits are larger when they allocate credit to prospective firms. Hence, incentives for banks to open new branches are high, explaining a lower value for ϕ. In Appendix A.10, we show that the (CC) locus crosses the (FC) only once.

Finally, some useful variables can be computed. The steady-state mass of active entrepreneurs is

$$e = \frac{1}{1 + n^* + \lambda/p(\phi)}, \tag{28}$$

the average mass of customers per bank is[20]

$$\tilde{M} = \left[\frac{c}{\Delta C'(1) - C(1)}\right]^{\frac{1}{\alpha}} \frac{\tilde{\varphi}^{\frac{1}{\alpha-1}}}{\varphi^{*\frac{1}{\alpha(\alpha-1)}}}, \tag{29}$$

the mass of banks is given by

$$b = \frac{e}{\tilde{M}} \tag{30}$$

and, finally, aggregate output can be written as

$$Y = eg(n^*). \tag{31}$$

Appendix A.8 shows in detail how to obtain equation (28) and Appendix A.9 provides the derivation of equation (29). Equation (30) simply states that the aggregate mass of banks is equal to the aggregate mass of customers in the economy (i.e., e) divided by the average mass of customers per bank, while equation (31) indicates that aggregate output is simply the product of output per firm multiplied by the mass of firms.

Given that the values of φ^*, ϕ and n^* exist and are unique, the next proposition follows:

Proposition 4. *The equilibrium exists and is unique.*

Proof. See Appendix A.10 □

[20] $\tilde{M}$ is the average mass of customers per bank and also the mass of customers of a bank with efficiency $\tilde{\varphi}$.

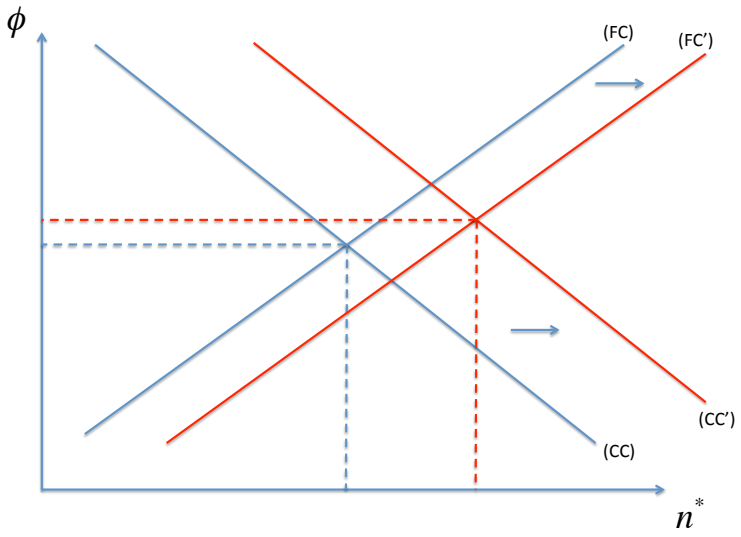

Figure 4: The impact on the (CC) and (FC) loci of a lower Δ.

3 The macroeconomic impact of overconcentration by banks

In this section we compare the equilibrium allocations of two economies. The first economy is one in which banks consider the value of ρ as given when they decide on K. This means that the derivative $\rho'(M)$ is absent in the equilibrium conditions (14) and (15). One can show that, in this economy, the resulting equilibrium conditions would be the same as in Section 2.7, with the difference that now $\Delta = 1$. The second economy is the one described in Section 2: the value taken by Δ is lower, as shown by equation (19). The comparison between these two economies allows us to understand the macroeconomic impact of the scale inefficiency in the model. This can be achieved by doing the comparative static with respect to Δ, taking into account the set of equilibrium conditions of Section 2.7.

Notice first that the value of Δ does not affect the (FE) and (ZCP) conditions. Hence, the threshold φ^* is independent of Δ. However, the value taken by ς decreases with Δ, as suggested by equation (27). Consequently, Δ affects the (CC) and (FC) loci through its effect on ς. It is easy to see that a lower Δ shifts the two loci to the right. The equilibrium value of n^* is thus higher in the second economy. The impact on ϕ is *a priori* ambiguous, but we show in Appendix A.11 that ϕ increases with a decrease

in Δ. This comparative static is illustrated on Figure 4. The blue lines are the (CC) and (FC) curves of the first economy, while the red lines characterize the situation in the second economy.

The following proposition summarizes these comparative statics and additionally describes the impact on average bank size, the mass of active entrepreneurs, the mass of banks and aggregate output:

Proposition 5. *Consider the set of equilibrium conditions* (CC), (FC), (FE), (ZCP), (17), (27), (28), (29), (30) *and* (31) *identifying* ϕ, n^*, φ^*, $\tilde{B}$, ς, ρ, e, $\tilde{M}$, b *and* Y. *A lower value of* Δ *implies higher values for* ϕ, n^*, ρ, $\tilde{M}$ *and* ς, *and lower values for* e *and* b. φ^* *and* $\tilde{B}$ *are independent of* Δ. Y *can either increase or decrease when* Δ *is lower.*

Proof. See Appendix A.11. □

Proposition 5 shows bank concentration increases when Δ decreases, in the sense that banks are larger (higher $\tilde{M}$) and the mass of banks is lower (lower b). The overlending behavior obviously increases average bank size and, because banks operate at an inefficiently large scale, a lower mass of banks survives.

Interestingly, Proposition 5 shows that the overlending behavior by banks also induces more concentration in the goods market (lower e and higher n^*). The intuition for this general-equilibrium effect is the following. Because overlending increases the cost of credit in the economy, it is natural that less firms enter this market. Moreover, since it is more costly to create new firms in the economy, firms' profits need to be larger in equilibrium for new entrepreneurs to be willing to enter the market. This can be achieved if firm size increases in equilibrium. Furthermore, as entrepreneurs prefer to work in the labor market, labor supply is larger in the economy with a lower Δ. This depresses the equilibrium wage rate and gives incentives for incumbent firms to increase their size.[21]

Finally, the impact of the scale inefficiency on aggregate output is ambiguous. On the one hand, it produces more concentration on the goods market. This negatively affects aggregate output because there are decreasing returns to scale in the goods sector: more can be produced when there are many small firms than when there are few large firms. On the other hand, when the scale inefficiency is present, some entrepreneurs prefer to be workers: this contributes positively to output since entrepreneurs are idle workers within the firm.

[21] Although it is easy to understand why the supply of entrepreneurs e is lower in the economy with a lower Δ, the result of a higher ϕ is less obvious as there are also fewer banks. The reason behind this result is related to the response of the wage rate, which is the opportunity cost of creating a firm. The change in the wage rate makes the supply of entrepreneurs less elastic than the supply of banks.

4 Welfare

We now consider the problem of a social planner who allocates credit to firms in a context where search frictions are given and the efficiency of each individual bank cannot be chosen. The solution to this problem is given by the following proposition:

Proposition 6. *The constrained-efficient allocations are a set of firm size n^*, a tightness ϕ, a bank efficiency threshold φ^*, an average bank value $\tilde{B}$, a common marginal agency cost σ and a mass of active entrepreneurs e such that the following conditions hold:*

$$\frac{g'(n^*)}{p(\phi)} = (1-\chi(\phi))\left[\frac{\pi(n^*)-\sigma}{r+\lambda} - \kappa\right], \tag{32}$$

$$\frac{\eta}{q(\phi)} = \chi(\phi)\left[\frac{\pi(n^*)-\sigma}{r+\lambda} - \kappa\right] \tag{33}$$

and

$$\sigma = \frac{C'(1)c^{\frac{\alpha-1}{\alpha}}}{[C'(1)-C(1)]^{\frac{\alpha-1}{\alpha}}}\varphi^{*-\frac{1}{\alpha}}. \tag{34}$$

together with conditions (FE)*,* (ZCP) *and* (28)*.*

Proof. See Appendix A.12. □

Three standard inefficiencies characterize the decentralized equilibrium. First, according to the scale inefficiency described earlier, banks create too many branches. Second, congestion externalities *à la* Hosios (1990) are not internalized in the negotiation process between a bank and an entrepreneur. Depending on how the bargaining power β compares with the elasticity of the matching function $\chi(\phi) = -\frac{p'(\phi)}{p(\phi)}\phi$, there may be too many (if $\beta > \chi(\phi)$) or too few branches (if $\beta < \chi(\phi)$) in the credit market. Third, there is a hold-up problem on the credit market as in Acemoglu and Shimer (1999): because the payment of the κ cost by the bank is sunk and there is continuous renegotiation between a bank and an entrepreneur, this induces banks to create too few branches.

The first inefficiency can be identified by comparing (34) with (27): one needs Δ to be equal to one for (27) to be identical to (34).

It is easier to identify the second inefficiency by considering an economy where the rule (10) is always the one considered instead of (11) to determine ρ. In this case, conditions (CC) and (FC) would read as

$$\frac{g'(n^*)}{p(\phi)} = (1-\beta)\left[\frac{\pi(n^*)-\sigma}{r+\lambda} - \kappa\right] \tag{35}$$

and

$$\frac{\eta}{q(\phi)} = \beta \left[\frac{\pi(n^*) - \sigma}{r + \lambda} - \kappa \right]. \tag{36}$$

By focusing on conditions (35) and (36) instead of (CC) and (FC), we can forget about the third inefficiency. Comparison of (32) and (33) with (35) and (36) reveals that congestion externalities are internalized if $\beta = \chi(\phi)$.

Finally, the last inefficiency can be identified by comparing (35) and (36) with (CC) and (FC). This shows that the two sets of conditions are identical if $\theta = 1$, confirming the presence of a hold-up problem in the credit market.

5 Quantitative analysis

5.1 Calibration

Table 1: Calibration: parameter values

Parameter	Description	Value
β	Bank's bargaining power	0.0875
α	Agency cost function convexity	1.1182
ε	Pareto distribution shape	9.4535
φ_0	Pareto distribution lower bound	1
c	Bank fixed operating cost	0.0125
ν	Bank entry cost	1
η	Branch opportunity cost	0.2593
κ	Firm set-up cost	10.7430
θ	Hold-up parameter	1
m_0	Matching function scale parameter	9.6879
χ	Matching function elasticity	0.5
r	Discount rate	0.04
λ	Firm death rate	0.0602
γ	Labor income share	2/3

We consider that a unit interval of time represents a year. Our calibration considers the US economy in 2014. To calibrate our model, we need to specify some functional forms. As is standard in the credit search literature (e.g. Wasmer and Weil (2004), Petrosky-Nadeau and Wasmer (2013)), we consider a Cobb-Douglas specification for the matching function:

$$m(\mathcal{E}, \mathcal{K}) = m_0 \mathcal{E}^{1-\chi} \mathcal{K}^{\chi}. \tag{37}$$

We also consider a Cobb-Douglas form for the production function:

$$g(n) = n^{\gamma}, \tag{38}$$

implying that the labor income share in aggregate output is equal to γ. We fix $\gamma = \frac{2}{3}$, as is standard in the RBC literature.

We assume a Pareto distribution for F with lower bound φ_0 and shape parameter $\varepsilon > \frac{1}{\alpha - 1}$:

$$F(\varphi) = 1 - \left(\frac{\varphi_0}{\varphi} \right)^{\varepsilon}. \tag{39}$$

The shape parameter is an index of the dispersion of productivity draws: dispersion decreases as ε increases, and the productivity draws are increasingly concentrated toward the lower bound φ_0.

We use data from the Federal Deposit Insurance Corporation (FDIC) on the distribution of branches across banks to calibrate several parameters of the benchmark economy. The Summary of Deposits Survey is filled annually by all FDIC-insured institutions. We focus on the 2014 cross-section of commercial banks and compute the Gini coefficient and the average number of branches among these institutions.[22]

The distribution of branches in the U.S. is fairly concentrated, as the Gini coefficient is 0.81.[23] This moment allows us to calibrate the curvature of the agency cost function α and the shape parameter ε. To see this, notice that concentration is high when α is low because banks with high efficiency can increase their size at a lower marginal cost. Moreover, as the dispersion of efficiencies across banks increases, concentration naturally increases too.

The average number of branches in the data is 15.03. This moment influences the calibration of α, the bargaining power β (mostly through the scale inefficiency in the

[22] The empirical literature on banking typically relies on other measures of concentration rather than the Gini coefficient, such as the Herfindahl-Hirschman or the C4 indexes. The advantage of these alternative measures is that they are also influenced by the number of banks in the industry, which is a relevant dimension if one wants to study the degree of competition in the sector, while the Gini coefficient is independent of this variable. As an illustration, the Gini coefficient has not changed much over the last twenty years (going from 0.70 to 0.81), while Figure 1 clearly shows an important decrease in the number of banks over this period, together with an important increase in the number of offices per bank. However, the number of banks is not discrete in our model and a measure such as the Herfindahl-Hirschman index would not be independent of the size of the economy assumed in the model. We thus choose to consider the Gini coefficient together with the average number of branches per bank as a better alternative.

[23] It is, though, less concentrated than financial variables such as financial assets (Gini = 0.92), loans (Gini = 0.90) and deposits (Gini = 0.90).

model) and the fixed cost parameter c, since it identifies average bank size.[24]

In order to assess the average efficiency in the banking sector, the empirical literature on banking computes a measure called *X-efficiency*. X-efficiency is calculated by estimating first a bank-specific multiplicative factor in the cost function, and then computing the ratio of this factor in the most efficient bank to the factor in each bank. We rely on the available estimates of X-efficiency to discipline the model. Specifically, we target an average X-efficiency ratio of 85.59%, in line with Evanoff and Ors (2008).[25,26] This moment plays a similar role as the Gini coefficient in our calibration, as it mainly helps us identify the shape parameter ε and the curvature of the agency cost function α.

As noticed by Petrosky-Nadeau and Wasmer (2013), the lack of data on the average search duration by banks and fund raisers renders the task of calibrating the scale parameter of the matching function m_0 quite difficult. We thus simply choose to target a duration for entrepreneurs of four months and then document results for other targeted durations in the sensitivity analysis of Section 5.4. Similarly, because we do not know which value to assign to the elasticity of the matching function, we simply set it equal to the commonly accepted value in the search literature of 0.5 and then recalibrate the model with other values in Section 5.4.

An important parameter for our quantitative exercise is θ, the parameter determining the importance of the hold-up problem. It turns out that, for small deviations of θ away from one, it is not possible to calibrate the model to match the targets described before. In particular, we show in Section 5.4 that one needs to accept search durations for fund raisers of several years to accommodate low values for θ: intuitively, banks are less willing to supply funds to entrepreneurs because they anticipate these will be appropriated by entrepreneurs, generating a long queue for credits. We thus choose to fix $\theta = 1$ and then document the implausibility of low values for θ in the section devoted to the sensitivity analysis.

To calibrate the remaining parameters, we proceed as follows. To identify κ and β, we target a loan rate of 12%, in line with Asea and Blomberg (1998), together with a firm size of 17 employees as in Guner, Ventura, and Xu (2008). The discount rate r is fixed at 4%. The firm death rate λ is fixed at a 6.02% annual rate as in Janiak and Santos Monteiro (2011). Without loss of generality, we normalize the bank entry cost and the lower bound of the Pareto distribution to 1. We fix the parameter η equal to

[24] In several credit search models in the literature, it is difficult to identify the bargaining power in the calibration. Because of the presence of the scale inefficiency, this is not the case in our model.

[25] Since Berger (1995) takes out the 1% most efficient banks in his sample, we proceed equally.

[26] As documented in Berger and Humphrey (1997), this is a common estimate for the US economy when parametric techniques are applied. Non-parametric techniques typically yield estimates around 72%.

Table 2: Concentration of branches: model versus data

Percentile	Data	Model
10%	0.67%	1.10%
50%	6.01%	7.018%
75%	13.90%	13.54%
90%	23.41%	21.47%
95%	29.63%	26.98%
99%	43.89%	38.33%

the equilibrium wage in the economy. This implies that we interpret each branch of a bank as a worker who spends time searching for a customer.

The targeted moments are accurately matched in our calibration exercise. The resulting parameter values are reported in Table 1. Appendix B explains in detail how we compute moments for the calibration and the quantitative exercise below.

5.2 Properties of the calibrated economy

In what follows, we describe some moments of the benchmark economy, which are not *a priori* restricted by our calibration strategy.

5.2.1 Branch distribution

As discussed before, we use the Gini coefficient of the distribution of branches as a target in our calibration strategy. Despite the fact that this moment is matched accurately, we need to check whether the model approximates reasonably well the whole distribution of branches across banks.[27] Table 2 shows the percentage of branches held by banks at different percentiles of the distribution and compares it with the data. Although the model does generate more branch concentration across banks than there is in the data,[28] the differences are reasonably small. In sum, the model does a good job at matching the entire distribution of branches across banks for the U.S.

[27]Notice that this may not be the case. For example, in the heterogenous agents Aiyagari-Bewley-Huggett literature, most existing models have a hard time obtaining the share of assets held by the one percent richer households observed in the data, even though the Gini coefficient is typically used as a calibration target (see e.g. Domeij and Heathcote (2004)).

[28]For example, the one percent largest banks in the model hold 61.7% of branches, while they hold 56.1% in the data.

5.2.2 Economies of scale

Many studies in the empirical literature on banking estimate the importance of economies of scale for this sector. The basic methodology regresses the logarithm of total costs on the logarithm of bank size for a cross section of banks. If the value of the estimated coefficient lies below one, then one interprets this result as evidence of economies of scale, while diseconomies of scale are found when the coefficient is larger than one. The basic methodology hardly finds evidence of economies of scale because the estimated coefficient is typically not significantly different from one. However, more advanced studies that incorporate in their models the ability of large banks to diversify risk do find evidence of scale economies (see e.g. Hughes and Mester (2013) or Wheelock and Wilson (2012), among others).

Our model does not consider aspects related to risk, but it does generate a distribution of banks of different sizes that allows to run a simple regression of the logarithm of bank costs on the logarithm of bank size and compare the results with the available evidence. In Appendix B.3 we calculate the value of the coefficient that one would estimate out of the cross section of banks of the calibrated economy. The value for this coefficient is 0.99, which is consistent with the evidence of moderate scale economies discussed earlier.

It may seem surprising to find evidence of small scale economies among the banks of our calibrated economy. Indeed, the overbranching behavior of banks places them to the right of the minimum of their average cost curve, which suggests the presence of diseconomies of scale. The reason for this result lies on the fact that, following the basic methodology, we compare costs across a cross-section of banks and do not compare observations of a single bank across time. Size is an endogenous variable in our cross section of banks: it is the unobserved heterogeneity in φ that generates the dispersion in bank size in our model. Hence, even though one could interpret the results as evidence of scale economies, it is perfectly consistent with the overbranching behavior of banks.

In Appendix B.3, we show that this a general result. In the context of our model, the equilibrium value of the cost of a bank is a linear function of its size M, where M corresponds to the size *chosen* by the bank and where the constant in this linear function is the fixed-operating cost c.[29] This linear property suggests that, by comparing a given bank with another bank that is one percent larger, total costs in the latter are less than one percent higher than in the former bank.

[29] See Appendix B.3.

5.2.3 Scale efficiency

Another measure of the efficiency of the banking sector that is sometimes estimated in the literature is the *scale efficiency*. It is obtained first by estimating a U-shaped average cost function for each bank in a sample. One then identifies the optimal input mix at the minimum of the curve and calculates the scale efficiency as the ratio of the average cost at the minimum to the actual average cost faced by the bank.

In our model, banks are not at the minimum of their average cost curve because of the incentives to allocate too much credit and the requirement to pay the entry cost ν. In Appendix B.5, we show how to obtain a measure of scale efficiency for our calibrated economy. In particular, we obtain a ratio of 87.3%. Unfortunately, scale efficiency is a measure that was mostly estimated in the nineties and there does not seem to be any recent study to which the 87.3% figure can be compared: for example, the work by Berger (1995) reports an estimate of 81.5% for the eighties. Our measure is thus slightly above the one of Berger (1995). Nevertheless, it is reasonable to suspect that the mergers and acquisitions in the US banking sector that occurred over the last twenty years would have increased this value. Empirical studies on the impact of mergers and acquisitions have shown little effect over scale efficiency. The paper by Peristiani (1997) documents small effects of mergers for the banks in the 80s, while the average estimates of scale efficiency in Al-Sharkas, Hassan, and Lawrence (2008) are similar whether one considers the period 1999 or 1987, though merged banks are on average slightly more efficient than non-merged banks.[30]

5.2.4 A calibration without the scale inefficiency

A possible validation of our model is to analyze how the calibrated economy would perform in a context without the scale inefficiency. In this section, we document that the scale inefficiency is required in the model in order to produce a realistic measure of X-efficiency. To do this, we set $\Delta = 1$ and calibrate the model in order to match all previous targets except for X-efficiency. Because we have one target less in the calibration, we set β to a particular value.[31] We perform this exercise over a grid for β, and compute the value of the (untargeted) measure of X-efficiency obtained from the calibrated model. We also compute the measure of scale efficiency previously discussed. The results of this exercise are shown in Table 3.

[30] We cannot use the estimated scale efficiency in Al-Sharkas, Hassan, and Lawrence (2008) as they rescale their index so that it attains values between zero and one for the least and the most efficient units in the sample.

[31] Setting β to a particular value is useful because this parameter has a lower and upper bound defined within the model. Leaving other parameters out of the calibration procedure yields similar results, which we do not reproduce here for the sake of brevity, but are available upon request.

Table 3: Calibration of an economy without scale inefficiency

β	X-efficiency	Scale efficiency
0.03	0.9778	0.9856
0.05	0.9778	0.9856
0.0875	0.9779	0.9856
0.15	0.9779	0.9856
0.3	0.9780	0.9854
0.5	0.9783	0.9850
0.7	0.9788	0.9845

For all values of β considered, the model can be calibrated up to a very high degree of precision; that is, the values of the loan rate, firm size, search duration for firms, average mass of branches per bank and the Gini coefficient of the distribution of branches in the model are all equal to their counterparts in the data. However, Table 3 shows that the model yields a value for X-efficiency that is very close to one, and far from the value found in empirical studies, and the same holds for the measure of scale efficiency. The economy without the scale inefficiency is thus characterized by a dispersion of efficiencies across banks that is too low as compared to the available evidence. We conclude that one needs the scale inefficiency to be present in the model for the calibration to deliver results in line with empirical studies of bank concentration.

5.3 Quantitative importance of the scale inefficiency

We now assess the quantitative impact of the scale inefficiency in our calibrated economy by considering an alternative economy where banks take the value of ρ as given in their decision process. In this alternative economy, the variable Δ is simply set equal to one while keeping the rest of the parameters as in the benchmark economy.

Table 4 reports statistics that allow to compare the two economies in steady state. The second column represents the calibrated economy (the economy with the scale inefficiency), while the third column is the alternative economy. We compare firm size, the loan rate, the equilibrium wage, search duration for firms and banks, the total mass of firms and banks, aggregate output and welfare.[32] Firm size, the loan rate and search duration for firms in the left column are moments we use in the calibration, while output, the mass of firms and banks, the equilibrium wage and welfare are normalized

[32]Welfare is output net of search, entry and agency costs, and it is computed as the flow welfare in expression (95) in Appendix A.12.3.

Table 4: The impact of the scale inefficiency

	Scale inefficiency		Constrained-eff. allocation
	included	excluded	
Loan rate	0.12	0.108	n.a.
Wage*	1	1.029	n.a.
Firm size	17.0	15.59	15.55
Mass of firms*	1	1.085	1.089
Average mass of branches per bank	15.03	1.55	1.55
Mass of banks*	1	10.50	10.54
Search duration for firms	0.333	0.329	0.101
Search duration for banks	0.032	0.032	0.105
Aggregate output*	1	1.024	1.026
Welfare*	1	1.047	1.048

Note: * variables normalized to one in the calibrated economy.

to one for this economy.

We see that, by removing the scale inefficiency, the loan rate becomes lower. This is intuitive: banks do not overlend in the alternative economy as a way to increase the repayment flow ρ, so this must be associated with a lower loan rate. Notice that the impact is quantitatively relevant, as the difference between the rates in the two economies is 1.2%.

The impact on bank concentration is illustrated by the effect on average bank size and the aggregate mass of banks in the economy. Our objective is not to replicate the overall increase in bank size observed on Figure 1 over the last fifty years, but it may be useful to compare these figures with the ones from columns 2 and 3. Interestingly, it turns out that, by removing the scale inefficiency, the decrease in average bank size (from 15.0 to 1.6) is similar to the one observed in the data since the bank deregulations of the seventies. The impact on the aggregate mass of banks is stronger though, as the ratio in the data is about 3 while it is 10 in the model. Nevertheless, if one takes into account that US GDP has been multiplied by a factor of about 3.5 according to NIPA tables over the last fifty years, then the impact on the mass of banks may be seen as similar.

Because the loan rate in the alternative economy is lower, this enhances firm creation. Agents reallocate from working on the labor market to fund raising. This decrease in labor supply increases the equilibrium wage by 2.9%. As a result, incumbent firms reduce their size from 17 employees to 15.6 and the mass of firms increases

by 8.5%.

Even though firms are smaller in the alternative economy, aggregate output is 2.4% larger. This is because there are more firms. Hence, the lower bank concentration also generates lower concentration on the goods market. The positive impact on aggregate output is due to the decreasing returns of the production function: more can be produced if workers are distributed among many firms since their marginal productivity becomes larger.

The table also shows that welfare increases by 4.7% once the scale inefficiency is excluded. To understand the quantitative importance of this inefficiency, we also report on the last column of the table the statistics for the constrained-efficient allocations. In this case, both the scale inefficiency and the congestion externalities are omitted. One can see that the numbers in this last column are quite similar to the case where the scale inefficiency is excluded (except for the durations). Congestion externalities are thus small from a welfare perspective: the welfare losses are mostly due to the scale inefficiency.

5.4 Sensitivity analysis

In this section, we investigate how the results in Table 4 change when we alter some of the targets used for the calibration. In particular, we report results for those parameters for which the identification is more debatable, such as the elasticity of the matching function, the stringency of the holdup problem and the target for X-efficiency. In Appendix C, we also report results for different targeted search durations for firms, and show that the results are similar to the benchmark case.

5.4.1 Elasticity of the matching function

In standard search and matching models, the distance between the elasticity of the matching function and the bargaining power of sellers determines the size of the inefficiencies resulting from congestion on the market. Constrained efficiency is restored when the so-called Hosios-Pissarides rule applies, i.e when those two parameters are equal.[33] The size of this elasticity also matters in the context of our model, as explained in Section 4.

As discussed in Section 5, we do not have a precise way to calibrate the elasticity of the matching function χ. In the benchmark calibration, χ is set equal to 0.5, in agreement with the search literature. In this section, we recalibrate the model

[33] See Hosios (1990) and Pissarides (2000).

Table 5: Sensitivity analysis: χ

χ		Scale inefficiency included	Scale inefficiency excluded	Constrained-eff. allocation
0.1	Loan rate	0.120	0.108	n.a.
	Search duration for firms	0.333	0.332	0.328
	Search duration for banks	0.032	0.033	0.038
	Aggregate output*	1	1.024	1.024
	Welfare*	1	1.047	1.047
0.3	Loan rate	0.120	0.108	n.a.
	Search duration for firms	0.333	0.331	0.211
	Search duration for banks	0.032	0.033	0.093
	Aggregate output*	1	1.024	1.025
	Welfare*	1	1.047	1.048
0.5	Loan rate	0.120	0.108	n.a.
	Search duration for firms	0.333	0.329	0.102
	Search duration for banks	0.032	0.032	0.105
	Aggregate output*	1	1.024	1.026
	Welfare*	1	1.047	1.048
0.7	Loan rate	0.120	0.108	n.a.
	Search duration for firms	0.333	0.327	0.035
	Search duration for banks	0.032	0.032	0.084
	Aggregate output*	1	1.024	1.026
	Welfare*	1	1.047	1.049
0.9	Loan rate	0.120	0.108	n.a.
	Search duration for firms	0.333	0.324	0.005
	Search duration for banks	0.032	0.032	0.051
	Aggregate output*	1	1.0241	1.027
	Welfare*	1	1.047	1.049

Note: * variables normalized to one in the economy without the scale inefficiency.

considering alternative values for χ and report results for the main variables in the model.

Table 5 shows the main results of the exercise, for alternative values of χ from 0.1 to 0.9.[34] It is evident from the fourth column of the table that the impact of the scale inefficiency is not very sensitive to the value assigned to the elasticity of the matching function. As expected, the impact on welfare is slightly larger for large values of the elasticity given that the gap with respect to the bargaining power is larger, but the values are very similar across all specifications.[35]

The last column of Table 5 shows the results when we exclude the scale inefficiency and the congestion externality. Once again, the elasticity of the matching function does not seem to change the results in a substantial manner. The sole exceptions are search duration for firms and for banks. A higher value for χ translates into a lower search duration for firms and a higher search duration for banks. The reason for this is intuitive. A high value for χ means that the market is congested on the firms' side, implying that the search duration for them should be lower in the constrained-efficient equilibrium.

5.4.2 Stringency of the hold-up problem

In this section we study the quantitative effects of varying the stringency of the hold-up problem, captured by parameter θ. In the benchmark exercise we set $\theta = 1$ and calibrate the model to obtain a search duration for firms of 1/3. However, when considering a value for θ lower than 1—thus increasing the stringency of the hold-up problem—a striking conclusion arises. For some small deviations of θ away from 1, the model cannot deliver a search duration for firms of 1/3 together with the rest of the targets discussed before. We interpret this result as evidence that $\theta \approx 1$.

To illustrate the difficulty of delivering reasonable targets when $\theta < 1$, we choose to target instead the search duration for banks to be equal to the value obtained in the benchmark calibration (around 12 days) when we recalibrate the model for other values of θ. Table 6 shows the results.

As the table shows, the calibrated model economy yields a search duration for firms of three and a half years when $\theta = 0.99$, and as high as 36 years when $\theta = 0.1$. This result is a direct consequence of the hold-up problem that arises when $\theta < 1$. When banks cannot recover the sunk cost κ in full, should the renegotiation process with

[34] Given space constraints, we do not report the results for all the variables in Table 4, but they are available upon request.

[35] The recalibrated value for β is the same across all specifications and identical to the benchmark (see Table 1). The only recalibrated parameter that changes across specifications is the scale parameter of the matching function m_0.

Table 6: Sensitivity analysis: θ

θ		Scale inefficiency		Constrained-eff. allocation
		included	excluded	
1	Loan rate	0.120	0.108	n.a.
	Search duration for firms	0.33	0.329	0.102
	Search duration for banks	0.032	0.032	0.105
	Aggregate output*	1	1.024	1.026
	Welfare*	1	1.047	1.048
0.99	Loan rate	0.120	0.108	n.a.
	Search duration for firms	3.561	3.474	0.330
	Search duration for banks	0.032	0.033	0.345
	Aggregate output*	1	1.022	1.048
	Welfare*	1	1.045	1.061
0.9	Loan rate	0.120	0.109	n.a.
	Search duration for firms	15.03	14.73	0.658
	Search duration for banks	0.032	0.033	0.731
	Aggregate output*	1	1.015	1.140
	Welfare*	1	0.037	1.120
0.7	Loan rate	0.120	0.110	n.a.
	Search duration for firms	24.856	24.47	0.821
	Search duration for banks	0.032	0.033	0.969
	Aggregate output*	1	1.011	1.229
	Welfare*	1	1.031	1.181
0.5	Loan rate	0.120	0.110	n.a.
	Search duration for firms	30.14	29.74	0.888
	Search duration for banks	0.032	0.032	1.086
	Aggregate output*	1	1.009	1.281
	Welfare*	1	1.027	1.219
0.1	Loan rate	0.120	0.111	n.a.
	Search duration for firms	36.39	36.01	0.974
	Search duration for banks	0.032	0.032	1.221
	Aggregate output*	1	1.007	1.345
	Welfare*	1	1.022	1.268

Note: * variables normalized to one in the economy without the scale inefficiency.

Table 7: Sensitivity analysis: X-efficiency

Targeted value for X-efficiency		Scale inefficiency included	Scale inefficiency excluded	Constrained-eff. allocation
0.8559	Loan rate	0.120	0.108	n.a.
	Aggregate output*	1	1.024	1.026
	Welfare*	1	1.047	1.048
0.8	Loan rate	0.120	0.101	n.a.
	Aggregate output*	1	1.038	1.040
	Welfare*	1	1.071	1.072
0.75	Loan rate	0.120	0.095	n.a.
	Aggregate output*	1	1.051	1.053
	Welfare*	1	1.092	1.093

Note: * variables normalized to one in the economy without the scale inefficiency.

the firm fail, banks reduce drastically the number of branches opened, thus increasing credit market tightness.

Notice that, when the hold-up problem is more stringent (i.e., θ is lower), the effect of removing the scale inefficiency over output and welfare becomes quantitatively less important. The reason for this is that the scale inefficiency is attenuated by the hold-up inefficiency when $\theta < 1$, implying that the output and welfare gains of removing the scale inefficiency are lower. Moreover, removing the scale inefficiency when $\theta < 1$ does not bring the economy too close to the constrained-efficient allocation (see the "Welfare" rows in columns 4 and 5 of Table 6). Indeed, Table 6 shows that when θ is low, correcting for all three inefficiencies present yields very large welfare effects, as large as 26.8% for $\theta = 0.1$ (most of the gains due to the hold-up problem being alleviated).

5.4.3 X-efficiency

In their review of the empirical literature trying to estimate X-efficiency in the banking sector, Berger and Humphrey (1997) document that results differ depending on the technique used for the estimation. We rely on parametric estimates to calibrate the benchmark economy, but non-parametric methodologies typically yield lower estimates of X-efficiency. Table 7 shows results when the targeted X-efficiency in the calibration is lower. One can see from the table that the impact is much larger in this case. For instance, for a targeted X-efficiency of 75%, output would increase by 5.1% and welfare by 9.2% had the scale inefficiency been removed. The reason for this is the following.

When X-efficiency is lower, one needs to recalibrate α and β to obtain the same Gini coefficient. In particular, one would need α to be larger and β lower, implying a larger scale inefficiency according to equation (19) (e.g. for a targeted X-efficiency of 75%, the calibrated values for α and β are 1.1819 and 0.0352 respectively).

6 Conclusions

We propose a model to study the macroeconomic impact of bank concentration. Entrepreneurs need financing to start up a business, which they obtain from banks. Branches and fund raisers meet in a credit market characterized by search frictions and agents can renegotiate the repayment flow paid by entrepreneurs to banks. In the game of renegotiation, banks have incentives to overbranch in order to negotiate higher repayment rates. Thus, banks operate at a too large scale. Moreover, the higher repayment rates charged by banks impact negatively on the decision of agents to become entrepreneurs, which results in fewer firms. Since there is a larger fraction of agents that decides to work, the wage rate decreases and firms become larger. To summarize, bank and firm concentration are too large compared to the efficient allocation.

We quantify this friction by calibrating the model to the US economy with data on the distribution of branches across banks and available estimates of X-efficiency in the banking sector. Our quantitative exercise suggests that the friction is sizable, and that it has substantial effects on output and welfare. A natural question that arises from our analysis is which policy would effectively reduce the inefficiency. It is straightforward to see that, in the context of our model, imposing a maximum interest rate would correct for the scale inefficiency. It would also be interesting to evaluate the consequences of fixing an exogenous limit on the number of branches that a given bank can open. The introduction of a limit on the number of branches in the current setup is not trivial given the heterogeneity in bank productivities: on the one hand, it may alleviate the scale inefficiency problem, but, on the other hand, it would harm high-productivity banks, which should efficiently have a large market share. We leave this exercise for future research.

References

Acemoglu, D., and W. B. Hawkins (2014): "Search with multi-worker firms," *Theoretical Economics*, 9(3), pp. 583–628.

Acemoglu, D., and R. Shimer (1999): "Holdups and Efficiency with Search Frictions," *International Economic Review*, 40(4), pp. 827–849.

Aiyagari, R., and S. Williamson (1999): "Credit in a Random Matching Model With Private Information," *Review of Economic Dynamics*, 2, pp. 36–64.

Al-Sharkas, A. A., M. K. Hassan, and S. Lawrence (2008): "The Impact of Mergers and Acquisitions on the Efficiency of the US Banking Industry: Further Evidence," *Journal of Business Finance and Accounting,*, 35(1), pp. 50–70.

Asea, P. K., and B. Blomberg (1998): "Lending cycles," *Journal of Econometrics*, 83, pp. 89–128.

Banerjee, A. V., and A. F. Newman (1993): "Occupational choice and the process of development," *Journal of Political Economy*, 101(2), pp. 274–298.

Bauducco, S., and A. Janiak (2015): "The impact of the minimum wage on capital accumulation and employment in a large-firm framework," Discussion paper.

Beck, T., A. Demirgüç-Kunt, L. Laeven, and R. Levine (2008): "Finance, Firm Size, and Growth," *Journal of Money, Credit and Banking*, 40(7), pp. 1379–1405.

Beck, T., A. Demirgüç-Kunt, and R. Levine (2007): "Bank concentration and fragility: impacts and mechanics," in *The Economics of Information and Uncertainty*, ed. by M. Carey, and R. M. Stulz, pp. 193–234. University of Chicago Press.

Berger, A. N. (1995): "The Profit-Structure Relationship in Banking–Tests of Market-Power and Efficient-Structure Hypotheses," *Journal of Money, Credit and Banking*, 27(2), pp. 404–431.

Berger, A. N., A. Demirgüç-Kunt, R. Levine, and J. G. Haubrich (2004): "Bank Concentration and Competition: An Evolution in the Making," *Journal of Money, Credit and Banking*, 36(3-2), pp. 433–451.

Berger, A. N., and D. B. Humphrey (1997): "Efficiency of financial institutions: International survey and directions for future research," *European Journal of Operation Research*, 98, pp. 175–212.

Berger, A. N., J. H. Leusner, and J. J. Mingo (1997): "The efficiency of bank branches," *Journal of Monetary Economics*, 40, pp. 141–162.

Bertola, G., and R. J. Caballero (1994): "Cross-Sectional Efficiency and Labour Hoarding in a Matching Model of Unemployment," *Review of Economic Studies*, 61(3), pp. 435–456.

Blanchard, O. J. (1985): "Debt, Deficits and Finite Horizons," *Journal of Political Economy*, 93, pp. 223–247.

Cahuc, P., F. Marque, and E. Wasmer (2008): "A theory of wages and labor demand with intrafirm bargaining and matching frictions," *International Economic Review*, 48(3), pp. 943–72.

Cao, M., and S. Shi (2001): "Screening, Bidding, and the Loan Market Tightness," *European Finance Review*, 5, pp. 21–61.

Cetorelli, N., and P. Strahan (2006): "Finance as a Barrier to Entry: Bank Competition and Industry Structure in Local U.S. Markets," *Journal of Finance*, 61(1), pp. 437–461.

Corbae, D., and P. D'Erasmo (2013): "A quantitative model of banking industry dynamics," Mimeo.

Corbae, D., and J. Ritter (2004): "Decentralized credit and monetary exchange without public record keeping," *Economic Theory*, 24, pp. 933–951.

De Elejalde, R. (2012): "Local entry decisions in the US banking industry," *ILADES working paper*.

Degryse, H., M. Kim, and S. Ongena (2009): *Microeconometrics of Banking: Methods, Applications and Results*. Oxford University Press.

den Haan, W. J., G. Ramey, and J. Watson (2003): "Liquidity flows and fragility of business enterprises," *Journal of Monetary Economics*, 50, pp. 1215–1241.

Diamond, D. W. (1984): "Financial intermediation and delegated monitoring," *Review of Economic Studies*, 51(3), pp. 393–414.

Diamond, D. W., and P. H. Dybvig (1983): "Bank runs, deposit insurance, and liquidity," *Journal of Political Economy*, 91(3), pp. 401–419.

Diamond, P. (1990): "Pairwise Credit in Search Equilibrium," *Quarterly Journal of Economics*, 105(2), pp. 285–319.

Domeij, D., and J. Heathcote (2004): "On the distributional effects of reducing capital taxes," *International Economic Review*, 45(2), pp. 523–554.

Evanoff, D. D., and E. Ors (2008): "The Competitive Dynamics of Geographic Deregulation in Banking: Implications for Productive Efficiency," *Journal of Money, Credit and Banking,*, 40(5), pp. 897–928.

Greenspan, A. (2010): "The crisis," *Brookings Papers on Economic Activity Spring*, pp. pp. 201–246.

Gu, C., F. Mattesini, C. Monnet, and R. Wright (2013): "Banking: a New Monetarist Approach," *Review of Economic Studies*, 80, pp. 636–662.

Guiso, L., P. Sapienza, and L. Zingales (2004): "Does local financial development matter?," *Quarterly Journal of Economics*, pp. pp. 929–969.

Guner, N., G. Ventura, and Y. Xu (2008): "Macroeconomic implications of size-dependent policies," *Review of Economic Dynamics*, 11, pp. 721–744.

Hopenhayn, H. A. (1992): "Entry, Exit, and firm Dynamics in Long Run Equilibrium," *Econometrica*, 60(5), pp. 1127–1150.

Hosios, A. J. (1990): "On the efficiency of matching and related models of search and unemployment," *Review of Economic Studies*, 57, pp. 279–298.

Hughes, J. P., and L. J. Mester (2013): "Who said large banks dont experience scale economies? Evidence from a risk-return-driven cost function," *Journal of Financial Intermediation*, 22, pp. 559–585.

Janiak, A. (2013): "Structural unemployment and the costs of firm entry and exit," *Labour Economics*, 23, pp. 1–19.

Janiak, A., and P. Santos Monteiro (2011): "Inflation and welfare in long-run equilibrium with firm dynamics," *Journal of Money, Credit and Banking*, 43(5), pp. 795–834.

Kehoe, T. J., and D. K. Levine (1993): "Debt-constrained asset markets," *Review of Economic Studies*, 60, pp. 865–888.

Kocherlakota, N. (1998): "Money is memory," *Journal of Economic Theory*, 81, pp. 232–251.

Kocherlakota, N., and N. Wallace (1998): "Incomplete record-keeping and optimal payment arragement," *Journal of Economic Theory*, 81, pp. 272–289.

Leland, H. E., and D. H. Pyle (1977): "Informational asymmetries, financial structure and financial intermediation," *Journal of Finance*, 32, pp. 371–387.

Lucas, R. E. (1978): "On the Size Distribution of Business Firms," *Bell Journal of Economics*, 9(2), pp. 508–523.

Martinez-Miera, D., and R. Repullo (2010): "Does competition reduce the risk of bank failure?," *Review of Financial Studies*, 23, 3638–3664.

Melitz, M. J. (2003): "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica*, 71(6), pp. 1695–1725.

Midrigan, V., and D. Y. Xu (2014): "Finance and Misallocation: Evidence from Plant-Level Data," *American Economic Review*, 104(2), pp. 422–458.

Nosal, E., and G. Rocheteau (2011): *Money, Payments, and Liquidity*. MIT Press.

Peristiani, S. (1997): "Do Mergers Improve the X-Efficiency and Scale Efficiency of U.S. Banks? Evidence from the 1980s," *Journal of Money, Credit and Banking*, 29(3), pp. 326–337.

Petrosky-Nadeau, N., and E. Wasmer (2013): "The Cyclical Volatility of Labor Markets under Frictional Financial Markets," *American Economic Journals: Macroeconomics*, 5(1), pp. 193–221.

Pissarides, C. A. (2000): *Equilibrium Unemployment Theory*. MIT Press.

Reiter, M. (2009): "Solving heterogeneous-agent models by projection and perturbation," *Journal of Economic Dynamics and Control*, 33(3), pp. 649–665.

Roberts, M. R., and A. Sufi (2009): "Renegotiation of financial contracts: Evidence from private credit agreements," *Journal of Financial of Economics*, 93, pp. 159–184.

Schaeck, K., M. Cihak, and S. Wolfe (2009): "Are more competitive banking systems more stable?," *Journal of Money, Credit and Banking*, 41, 567–807.

Sealy, C. W., and J. T. Lindley (1977): "Inputs, outputs, and a theory of production and cost at depository financial institutions," *Journal of Finance*, 32, pp. 1251–1266.

Shi, S. (2005): "The extent of the market and optimal specialization," *Economic Theory*, 25, pp. 333–351.

Smith, E. (1999): "Search, concave production, and optimal firm size," *Review of Economic Dynamics*, 2(2), pp. 456–471.

Stole, L. A., and J. Zwiebel (1996a): "Intra-Firm Bargaining under Non-Binding Contracts," *Review of Economic Studies*, 63(3), pp. 375–410.

——— (1996b): "Organizational Design and Technology Choice under Intrafirm Bargaining," *American Economic Review*, 86(1), pp. 195–222.

Wasmer, E., and P. Weil (2004): "The macroeconomics of labor and credit market imperfections," *American Economic Review*, 94(4), pp. 944–963.

Wheelock, D. C., and P. W. Wilson (2012): "Do Large Banks Have Lower Costs? New Estimates of Returns to Scale for U.S. Banks," *Journal of Money, Credit and Banking*, 44(1), pp. 171–199.

A Proofs

A.1 Repayment rate ρ

To show how to derive the solution (17) to (15), first consider the differential equation

$$\frac{\partial \rho}{\partial M}\frac{M}{\rho} = -\frac{1}{1-\beta}.$$

The solution is

$$\rho = \mathcal{C} M^{-\frac{1}{1-\beta}},$$

where $\mathcal{C}$ is a constant of integration.

Consider now the problem without a constant

$$\frac{\rho}{1-\beta} = C'_{\varphi}(M) - \frac{\partial \rho}{\partial M} M. \tag{40}$$

A guess for the solution is

$$\rho = \mathcal{C}(M) M^{-\frac{1}{1-\beta}}, \tag{41}$$

with derivative

$$\frac{d\rho}{dM} = \mathcal{C}'(M) M^{-\frac{1}{1-\beta}} - \frac{1}{1-\beta} M^{-\frac{1}{1-\beta}-1} \mathcal{C}(M). \tag{42}$$

By replacing (41) and (42) in (40), we obtain

$$\mathcal{C}'(M) = C'_{\varphi}(M) M^{\frac{\beta}{1-\beta}}. \tag{43}$$

By integrating this equation, we have

$$\mathcal{C}(M) = \int_0^M C'_{\varphi}(z) z^{\frac{\beta}{1-\beta}} dz + \mathcal{H} \tag{44}$$

where $\mathcal{H}$ is a constant of integration. With a change of variable $u = \frac{z}{M}$, it can be rewritten as

$$\mathcal{C}(M) = M^{\frac{1}{1-\beta}} \int_0^1 C'_{\varphi}(uM) u^{\frac{\beta}{1-\beta}} du + \mathcal{H}. \tag{45}$$

This last equation implies that the solution to (40) is

$$\rho = \int_0^1 u^{\frac{\beta}{1-\beta}} C'_{\varphi}(uM) du + \mathcal{H} M^{-\frac{1}{1-\beta}}. \tag{46}$$

As in Cahuc, Marque, and Wasmer (2008), we focus on a solution that implies that

$\lim_{M\to 0} M\rho = 0$. A necessary condition for this is $\mathcal{H} = 0$. Accordingly, the solution to the problem with constant (15) is (17) with

$$\Delta = \frac{\int_0^1 h(u)C'_\varphi(uM)du}{C'_\varphi(M)}, \tag{47}$$

an overlending factor, where $h(u) \equiv \frac{u^{\frac{\beta}{1-\beta}}}{1-\beta}$. The numerator in (47) is a weighted average of all inframarginal costs. Notice that the density $h(u)$ satisfies $\int_0^1 h(u)du = 1$. It assigns larger weights to inframarginal costs the larger the entrepreneur's bargaining power $(1-\beta)$ is.

Because C is homogenous of degree α, it follows that $C'(uM) = C'(M)u^{\alpha-1}$, implying that (47) can be rewritten as in (19).

A.2 First-order condition of the bank

To obtain (CC), first notice that we can rewrite $M\rho'(M)$ as

$$M\rho'(M) = (1-\Delta)C'_\varphi(M) \tag{48}$$

The proof is the following. Start by deriving (17):

$$M\rho'(M) = M\int_0^1 u^{\frac{1}{1-\beta}} C''_\varphi(uM)du.$$

Integration by part yields

$$M\rho'(M) = C'_\varphi(M) - \int_0^1 \frac{u^{\frac{\beta}{1-\beta}}}{1-\beta} C'_\varphi(uM)du$$

and

$$M\rho'(M) = C'_\varphi(M) - \Delta C'_\varphi(M),$$

yielding (48).

This property, together with (17), implies that the first-order condition (14) for the bank can be rewritten as in equation (CC).

Notice also that

$$\kappa + \frac{\eta}{\phi p(\phi)} = \frac{\rho - \varsigma}{r+\lambda}. \tag{49}$$

A.3 The value of an entering bank

Once a bank has entered the market, it opens a mass of branches sufficiently large such that its mass of customers jumps to its long-run value. This property has already been noticed in papers whose models share a similar structure to ours, such as Acemoglu and Hawkins (2014) and Janiak (2013). This can be seen easily seen from the first-order condition for K. Derive (8) with respect to K:

$$\frac{\eta}{\phi p(\phi)} + \kappa = \frac{\partial B(M';\varphi)}{\partial M'}. \tag{50}$$

The first-order condition above indicates that the future value of M' only depends on the credit-market tightness, which is constant because the economy is in steady state. Hence, M immediately jumps to its long-run value once the bank has entered the market.

Given this property of the model, we now show that the value of a bank upon entry can be written as in equation (20). To calculate this value, we need to calculate the value of an incumbent, as well as the cost that an entrant pays (including the required cost for M to jump to its long-run value).

$B(M(\varphi);\varphi)$ is the value of an established bank with efficiency φ which has reached its long-run size and $K(\varphi)$ the value of K chosen by an incumbent:

$$rB(M(\varphi);\varphi) = \rho M(\varphi) - \eta K - C_\varphi(M(\varphi)) - c - \kappa K(\varphi)\phi p(\phi).$$

From equation (9), we can establish that $K(\varphi) = M(\varphi)\frac{\lambda}{\phi p(\phi)}$. We have that

$$rB(M(\varphi);\varphi) = \rho M(\varphi) - \left[\frac{\eta}{\phi p(\phi)} + \kappa\right]\lambda M(\varphi) - C_\varphi(M(\varphi)) - c.$$

Using (49), we replace ρ in the equation above and get

$$rB(M(\varphi);\varphi) = r\left[\frac{\eta}{\phi p(\phi)} + \kappa\right] M(\varphi) + \Delta C'_\varphi(M(\varphi))M(\varphi) - C_\varphi(M(\varphi)) - c. \tag{51}$$

$B(0,\varphi)$ is the value of an entering bank with efficiency φ:

$$B(0,\varphi) = \frac{1}{1+rdt}\left\{B(M(\varphi);\varphi) - [\eta K_0(\varphi) + \kappa K_0(\varphi)\phi p(\phi) + c]dt\right\},$$

with $K_0(\varphi)$ the mass of branches opened upon entry.

From (9),

$$K_0(\varphi) = \frac{M(\varphi)}{\phi p(\phi) dt}.$$

By using this relation and letting $dt \to 0$, we obtain

$$B(0,\varphi) = B(M(\varphi);\varphi) - \left(\frac{\eta}{\phi p(\phi)} + \kappa\right) M(\varphi).$$

Finally, replacing (51) in the equation above yields (20).

A.4 Bank sizes

In this section we prove Lemma 1. The proof is based on Corollary 1, which establishes that all banks share the same ς.

Consider two firms with productivity φ_1 and φ_2 respectively, financing $M(\varphi_1)$ and $M(\varphi_2)$ firms each. We first want to show that

$$M(\varphi_2) = M(\varphi_1)\left(\frac{\varphi_2}{\varphi_1}\right)^{\frac{1}{\alpha-1}}.$$

The proof is the following. Given that banks all share the same ς:

$$\frac{C'(M(\varphi_1))}{\varphi_1} = \frac{C'(M(\varphi_2))}{\varphi_2}. \tag{52}$$

Given that C is homogenous of degree α, its derivative is homogenous of degree $\alpha - 1$. Hence,

$$\frac{C'(1)M(\varphi_1)^{\alpha-1}}{\varphi_1} = \frac{C'(1)M(\varphi_2)^{\alpha-1}}{\varphi_2}. \tag{53}$$

By rewriting this equation we obtain the property (22).

Now we show property (23). From (21), first show that

$$R(\varphi) = \frac{M(\varphi)^{\alpha}}{\varphi}\left(\Delta C'(1) - C(1)\right), \tag{54}$$

which is due to the fact that C is homogenous of degree α. Hence,

$$\frac{R(\varphi_2)}{R(\varphi_1)} = \frac{M(\varphi_2)^{\alpha}}{\varphi_2}\frac{\varphi_1}{M(\varphi_1)^{\alpha}}. \tag{55}$$

By using equation (22), we obtain (23).

A.5 Free entry condition

In this section we show how to obtain equation (FE). Free entry equals the sunk entry cost ν to expected profits in the banking sector:

$$\nu = [1 - F(\varphi^*)] \int_{\varphi^*}^{\infty} B(0, \varphi) \frac{dF(\varphi)}{1 - F(\varphi^*)}.$$

Notice that, in the equation above, a bank that chooses to leave the sector (the productivity of which is below φ^*) makes zero profits.

From (20), this equation can be rewritten as

$$r\nu = [1 - F(\varphi^*)] \int_{\varphi^*}^{\infty} (R(\varphi) - c) \frac{dF(\varphi)}{1 - F(\varphi^*)}.$$

Given (23) from Corollary 1,

$$r\nu = [1 - F(\varphi^*)] \int_{\varphi^*}^{\infty} \left(\left(\frac{\varphi}{\tilde{\varphi}} \right)^{\frac{1}{\alpha - 1}} R(\tilde{\varphi}) - c \right) \frac{dF(\varphi)}{1 - F(\varphi^*)},$$

$$r\nu = [1 - F(\varphi^*)] \left[\frac{R(\tilde{\varphi})}{\tilde{\varphi}^{\frac{1}{\alpha - 1}}} \int_{\varphi^*}^{\infty} \varphi^{\frac{1}{\alpha - 1}} \frac{dF(\varphi)}{1 - F(\varphi^*)} - c \right].$$

From the definition of aggregate efficiency in the banking sector (24):

$$r\nu = [1 - F(\varphi^*)] \left[R(\tilde{\varphi}) - c \right], \tag{56}$$

which gives (FE).

A.6 Bank exit

In this section we show how to obtain equation (ZCP). First calculate average flow profits in the sector

$$r\tilde{B} = \int_{\varphi^*}^{\infty} (R(\varphi) - c) \frac{dF(\varphi)}{1 - F(\varphi^*)}. \tag{57}$$

Given (23) from Corollary 1,

$$r\tilde{B} = \int_{\varphi^*}^{\infty} \left(\left(\frac{\varphi}{\varphi^*} \right)^{\frac{1}{\alpha - 1}} R(\varphi^*) - c \right) \frac{dF(\varphi)}{1 - F(\varphi^*)}, \tag{58}$$

$$r\tilde{B} = \frac{R(\varphi^*)}{\varphi^{*\frac{1}{\alpha - 1}}} \int_{\varphi^*}^{\infty} \varphi^{\frac{1}{\alpha - 1}} \frac{dF(\varphi)}{1 - F(\varphi^*)} - c. \tag{59}$$

From the definition of aggregate efficiency in the banking sector (24):

$$r\tilde{B} = R(\varphi^*)\left(\frac{\tilde{\varphi}}{\varphi^*}\right)^{\frac{1}{\alpha-1}} - c. \tag{60}$$

Given that a bank with productivity φ^* makes zero profits, it has to be that

$$R(\varphi^*) = c. \tag{61}$$

Replacing this equation in (60) produces (ZCP).

.

A.7 Aggregate inefficiency of the banking sector

In this section we show how to obtain equation (27). We first need to show that the amount of firms the least efficient bank finances is

$$M^* = \left[\frac{c\varphi^*}{\Delta C'(1) - C(1)}\right]^{\frac{1}{\alpha}}. \tag{62}$$

This can be shown as follows. The least efficient bank makes zero profits, meaning that $R^* = c$. Hence,

$$c\varphi^* = \Delta C'(M^*)M^* - C(M^*). \tag{63}$$

Given that C is homogenous of degree α:

$$c\varphi^* = M^{*\alpha}\left(\Delta C'(1) - C(1)\right). \tag{64}$$

By rewriting the equation above, one can get (62).

Equation (62) allows us to identify the measure of performance of the banking sector:

$$\varsigma = \frac{\Delta C'(1) M^{*\alpha-1}}{\varphi^*}, \tag{65}$$

$$\varsigma = \frac{\Delta C'(1)}{\varphi^*}\left[\frac{c\varphi^*}{\Delta C'(1) - C(1)}\right]^{\frac{\alpha-1}{\alpha}}. \tag{66}$$

By rewriting the equation above, one obtains equation (27).

A.8 Mass of active entrepreneurs

In steady state, it is the case that

$$\mathcal{E} = \frac{\lambda}{p(\phi)} e$$

by equating the flow of firms being created to the flow of firms being destroyed.

The mass of workers plus the mass of active entrepreneurs and the mass of fund raisers must add to one. Given that there is one active entrepreneur per firm as well as n^* workers, we have that

$$1 = \mathcal{E} + e(n^* + 1).$$

By combining the two equations above, one obtains equation (28).

A.9 Average bank size

In this section we show how to obtain equation (29).

Average bank size can be calculated as follows:

$$\tilde{M} = \int_{\varphi^*}^{\infty} M(\varphi) \frac{dF(\varphi)}{1 - F(\varphi^*)}.$$

Given relation (22) from Lemma 1, we can rewrite the equation above as

$$\tilde{M} = \int_{\varphi^*}^{\infty} M^* \left(\frac{\varphi}{\varphi^*} \right)^{\frac{1}{\alpha - 1}} \frac{dF(\varphi)}{1 - F(\varphi^*)},$$

or equivalently as

$$\tilde{M} = \frac{M^*}{{\varphi^*}^{\frac{1}{\alpha - 1}}} \int_{\varphi^*}^{\infty} \varphi^{\frac{1}{\alpha - 1}} \frac{dF(\varphi)}{1 - F(\varphi^*)}.$$

From the definition of aggregate efficiency in the banking sector (24), the equation above can be rewritten as

$$\tilde{M} = M^* \left(\frac{\tilde{\varphi}}{\varphi^*} \right)^{\frac{1}{\alpha - 1}}. \tag{67}$$

Finally, with equation (62), we rewrite (67) as

$$\tilde{M} = \left[\frac{c\varphi^*}{\Delta C'(1) - C(1)} \right]^{\frac{1}{\alpha}} \left(\frac{\tilde{\varphi}}{\varphi^*} \right)^{\frac{1}{\alpha - 1}},$$

which yields equation (29).

A.10 Existence and uniqueness of the equilibrium

The paper by Melitz (2003) shows that the (ZCP) locus cuts the (FE) locus once from above, implying that there exists an equilibrium value for φ^* that is unique. We now study the behavior of the (FC) and (CC) loci.

These two loci follow continuous paths given the assumptions on the functions m and g. The derivative of ϕ with respect to n^* along the (CC) locus is

$$\frac{d\phi}{dn^*}|_{(\text{CC})} = -\frac{\beta/\eta}{r+\lambda}\frac{q(\phi)^2}{q'(\phi)}\pi'(n^*). \tag{68}$$

Given that the parameters β, η, r and λ are strictly positive, that $\pi'(n^*) > 0$ and $q'(\phi) > 0$, the (CC) locus has a strictly negative slope (except should ϕ take the value zero, for which the slope would be null given that $q(0) = 0$).

From condition (CC), we can see that, when ϕ tends to infinity, n^* tends to a strictly positive value ω defined such as $\pi(\omega) = \varsigma + (r+\lambda)\frac{\kappa}{\beta}$, while $n^* \to \infty$ when $\phi \to 0$.

The derivative of ϕ with respect to n^* along the (FC) locus is

$$\frac{d\phi}{dn^*}|_{(\text{FC})} = \frac{g''(n^*)}{g'(n^*)}\frac{p(\phi)}{p'(\phi)} - \frac{1-\beta}{r+\lambda}\frac{\pi'(n^*)}{g'(n^*)}\frac{p(\phi)^2}{p'(\phi)}. \tag{69}$$

Given that $\beta \in (0,1)$, that the parameters η, r and λ are strictly positive, that $\pi'(n^*) > 0$, $g'(n^*) > 0$, $g''(n^*) < 0$ and $p'(\phi) < 0$, the (FC) locus has a strictly positive slope.

To understand the limits of the (FC) curve, first notice that (FC) can be rewritten as

$$p(\phi) = \frac{r+\lambda}{1-\beta}\frac{g'(n^*)}{\pi(n^*) - \varsigma - (r+\lambda)\theta\kappa}.$$

When, $n^* \to \infty$, it has to be that $\phi \to \infty$. Indeed, given that $g'(n^*) \to 0$ and $\pi(n^*) \to \infty$ when $n^* \to \infty$, $p(\phi) \to 0$ when $n^* \to \infty$, that is, $\phi \to \infty$ when $n^* \to \infty$.

Moreover, when $\phi \to 0$, it must be the case that n^* tends to a positive value x such that $\pi(x) = \varsigma$ ($g'(x)$ is strictly positive and finite in this case).

Hence, given the analysis above, it must be the case that the two loci cross only once: the equilibrium exists and is unique.

A.11 The macro impact of the scale inefficiency

Consider the set of equilibrium conditions (CC), (FC), (FE), (ZCP), (17), (27), (28), (29), (30) and (31) identifying ϕ, n^*, φ^*, $\tilde{B}$, ς, ρ, e, $\tilde{M}$, b and Y. From the discussion in Section 3, it is easy to understand why φ^* and $\tilde{B}$ are independent of Δ. We now

show in details the impact on the variables ς, n^*, ϕ, e, $\tilde{M}$, ρ, b and Y.

A.11.1 Aggregate inefficiency of the banking sector

Differentiating (27) with respect to Δ yields

$$\frac{d\varsigma}{d\Delta} = \varphi^{*-\frac{1}{\alpha}} c^{\frac{\alpha-1}{\alpha}} C'(1)[\Delta C'(1) - C(1)]^{\frac{1-\alpha}{\alpha}} \left[\frac{1/\alpha \Delta C'(1) - C(1)}{\Delta C'(1) - C(1)}\right].$$

Notice that $C(\cdot)$ is homogeneous of degree $\alpha > 1$. Then, applying Euler's homogeneous function theorem, $\alpha C(1) = C'(1)$. It follows that the numerator of the last term in square brackets is negative, since $\Delta < 1$. All remaining terms are positive, thus

$$\frac{d\varsigma}{d\Delta} < 0.$$

A.11.2 Firm size

Differentiating (FC) with respect to Δ yields

$$\left[\frac{g''(n^*)}{p(\phi)} - \frac{1-\beta}{r+\lambda}\pi'(n^*)\right]\frac{dn^*}{d\Delta} - g'(n^*)\frac{p'(\phi)}{p(\phi)^2}\frac{d\phi}{d\Delta} + \frac{1-\beta}{r+\lambda}\frac{d\varsigma}{d\Delta} = 0, \tag{70}$$

while by considering (CC), we have

$$-\beta\frac{\pi'(n^*)}{r+\lambda}\frac{dn^*}{d\Delta} - \eta\frac{q'(\phi)}{q(\phi)^2}\frac{d\phi}{d\Delta} + \frac{\beta}{r+\lambda}\frac{d\varsigma}{d\Delta} = 0. \tag{71}$$

By combining (70) and (71), we can make $\frac{d\phi}{d\Delta}$ disappear:

$$\begin{aligned}\left[\frac{g''(n^*)}{g'(n^*)}\frac{p(\phi)}{p'(\phi)} - \frac{1-\beta}{r+\lambda}\frac{\pi'(n^*)}{g'(n^*)}\frac{p(\phi)^2}{p'(\phi)}\right]\frac{dn^*}{d\Delta} + \frac{p(\phi)^2}{p'(\phi)g'(n^*)}\frac{1-\beta}{r+\lambda}\frac{d\varsigma}{d\Delta} \\ = -\frac{q(\phi)^2}{\eta q'(\phi)}\beta\frac{\pi'(n^*)}{r+\lambda}\frac{dn^*}{d\Delta} + \frac{q(\phi)^2}{\eta q'(\phi)}\frac{\beta}{r+\lambda}\frac{d\varsigma}{d\Delta}\end{aligned} \tag{72}$$

and express $\frac{dn^*}{d\Delta}$ directly as function of $\frac{d\varsigma}{d\Delta}$:

$$\frac{dn^*}{d\Delta} = \frac{\zeta(\phi, n^*)}{\frac{g''(n^*)}{g'(n^*)}\frac{p(\phi)}{p'(\phi)} + \zeta(\phi, n^*)}\frac{d\varsigma}{d\Delta} < 0. \tag{73}$$

with $\zeta(\phi, n^*) = \left(\frac{q(\phi)^2}{\eta q'(\phi)}\frac{\beta}{r+\lambda} - \frac{1-\beta}{r+\lambda}\frac{p(\phi)^2}{p'(\phi)g'(n^*)}\right)\pi'(n^*) > 0$.

A.11.3 Tightness

To show that ϕ increases when Δ decreases, rewrite (CC), (FC) as follows:

$$\frac{g'(n^*)}{p(\phi)(1-\beta)} = \frac{\pi(n^*)-\varsigma}{r+\lambda} - \theta\kappa \quad \text{and} \quad \frac{\kappa}{\beta} + \frac{\eta}{\phi p(\phi)\beta} = \frac{\pi(n^*)-\varsigma}{r+\lambda} - \theta\kappa.$$

By equating the left-hand sides of these two equations, we have that

$$g'(n^*) = \frac{1-\beta}{\beta}p(\phi)\kappa + \frac{1-\beta}{\beta}\frac{\eta}{\phi}.$$

Hence, given that n^* increases when Δ decreases and given that $g'(.)$ and $p(.)$ are decreasing functions, it has to be that ϕ increases when Δ decreases.

A.11.4 Active entrepreneurs

Given that both n^* and ϕ increase when Δ decreases, from equation (28), one can see that the mass of active entrepreneurs decreases when Δ through its impact on n^* and ϕ.

A.11.5 Repayment rate

Equation (17) shows that the equilibrium value for ρ is increasing in the equilibrium values for ς and n^*. Given that both ς and n^* increase when Δ decreases, it has to be that ρ increases when Δ decreases.

A.11.6 Average bank size

Equation (29) shows a relation between $\tilde{M}$, Δ and φ^*. Given that the equilibrium value of φ^* is independent of Δ, $\tilde{M}$ increases with a decrease in Δ (equation (29) shows a decreasing relation between $\tilde{M}$ and Δ).

A.11.7 Banks

Equation (30) shows that the equilibrium value for b is decreasing in $\tilde{M}$ and increasing in e. Given that, for a decrease in Δ, $\tilde{M}$ increases and e decreases, it must be the case that b decreases for an decrease in Δ.

A.11.8 Aggregate output

Differentiating equation (31) with respect to Δ yields

$$\frac{dY}{d\Delta} = \frac{Y}{n^*}[\gamma - en^*(1-\Psi)]\frac{dn^*}{d\Delta}, \tag{74}$$

where equation (28) has been used and γ is the elasticity of the production function with respect to n and $\Psi = \frac{\lambda p'(\phi)}{p(\phi)^2}\Gamma < 0$ and $\Gamma > 0$ is such that $\frac{d\phi}{d\Delta} = \Gamma\frac{dn^*}{d\Delta}$. It is easy to see that the sign of this derivative is ambiguous and depends on the sign of the expression in square brackets of equation (74).

A lower Δ has an ambiguous effect on output. To see why, notice that the derivative of Y with respect to Δ is

$$\frac{dY}{d\Delta} = g(n^*)\frac{de}{d\Delta} + eg'(n^*)\frac{dn^*}{d\Delta}, \tag{75}$$

$$= \frac{Y}{n^*}[\gamma - en^*(1-\Psi)]\frac{dn^*}{d\Delta}, \tag{76}$$

where γ is the elasticity of the production function with respect to n; $\Psi = \frac{\lambda p'(\phi)}{p(\phi)^2}\Gamma < 0$ and $\Gamma > 0$ is such that $\frac{d\phi}{d\Delta} = \Gamma\frac{dn^*}{d\Delta}$.

It is clear from expression (75) that how Y reacts to bank concentration depends on how n^* and e react, and the relative contribution of these to aggregate production. To gain intuition, assume first that the credit-market tightness ϕ does not react to changes in Δ. Then, $\Psi = 0$ in equation (76). As discussed before, a lower Δ always implies a higher n^*, so the last derivative is always negative. If $\gamma = 1$[36], the term in square brackets is positive. Intuitively, entrepreneurs are idle workers once production takes place and, consequently, they enter as a fixed factor of production in our model. Thus, if $\gamma = 1$, an economy with lower e will pay a lower fixed cost of production, and will be able to devote resources to productive labor n^*[37]. Notice that this effect is dampened if $\Psi < 0$, because the increase in credit-market tightness depresses e and increases the number of entrepreneurs that search funding.

Consider now the case in which $\Psi = 0$ and $\gamma < 1$. In this case, the term in square brackets can be either positive or negative, depending on the value of n^*. Since $\gamma < 1$, the economy will display an efficient scale of production $\tilde{n}$ where aggregate average costs are minimized. For $n^* < \tilde{n}$, an increase in n^* due to a decrease in Δ will increase Y, and the converse is true for $n^* > \tilde{n}$. If $\Psi < 0$, the positive effect on Y of the increase in n^* is dampened, thus rendering the second scenario more likely.

[36]Notice that a CRS production function displays $\gamma = 1$, while $\gamma < 1$ for a DRS production function.

[37]If $\gamma = 1$, $\phi \to \infty$ and entrepreneurs were also workers in their own firms, then $\frac{dY}{d\Delta} = 0$. In this case, the scale of production would not matter for Y. If, on the other hand, $\gamma < 1$, $\frac{dY}{d\Delta} > 0$ because of decreasing returns to scale in production: production is larger when it is carried out by a large set of small firms than by a small set of of large firms.

A.12 Constrained-efficient allocations

To understand the the nature of the constrained-efficient allocations, we first consider simpler versions of the model in Section 2. This is because the optimality conditions also hold in these alternative frameworks. In particular we simplify along two dimensions: bank heterogeneity and free entry and exit of banks.

A.12.1 Case with a fixed amount of homogenous banks

We start with the simplest case where both simplifications are considered. In this case, notice that we have $e = M$, where M is the mass of customers of the representative bank. The social planner problem can be written as follows:

$$\Omega(e) = \max_{n,\mathcal{K}} \frac{1}{1+rdt} \left\{ \left[eg(n) - \eta\mathcal{K} - C(e) \right] dt + \Omega(e') - \frac{\kappa p(\phi)\mathcal{E}dt}{1+rdt} \right\} \tag{77}$$

such that

$$e' = (1 - \lambda dt)e + p(\phi)\mathcal{E}dt$$

with

$$\phi = \frac{1 - e(n+1)}{\mathcal{K}}$$

and

$$\mathcal{E} = 1 - e(n+1)$$

To solve this problem, we start by obtaining the first-order condition for n. We derive problem (77) with respect to n and set the derivative to zero:

$$0 = \frac{1}{1+rdt} \left\{ eg'(n)dt + \Omega'(e')\frac{\partial e'}{\partial n} - \kappa \left(p(\phi)\frac{\partial \mathcal{E}}{\partial n} + p'(\phi)\frac{\partial \phi}{\partial n}\mathcal{E} \right) dt \right\}. \tag{78}$$

Given that

$$\frac{\partial e'}{\partial n} = \left(p(\phi)\frac{\partial \mathcal{E}}{\partial n} + p'(\phi)\frac{\partial \phi}{\partial n}\mathcal{E} \right) dt$$

we can simplify (78) as

$$0 = eg'(n) + \left[\Omega'(e') - \kappa \right] \left(p(\phi)\frac{\partial \mathcal{E}}{\partial n} + p'(\phi)\frac{\partial \phi}{\partial n}\mathcal{E} \right).$$

Then, notice that

$$\frac{\partial \mathcal{E}}{\partial n} = -e$$

and

$$\frac{\partial \phi}{\partial n} = -\frac{e}{\mathcal{K}}.$$

Hence,

$$0 = eg'(n) - \left[\Omega'(e') - \kappa\right]\left(p(\phi)e + p'(\phi)e\frac{\mathcal{E}}{\mathcal{K}}\right)$$

By rearranging terms in the equation above, we get

$$\frac{g'(n)}{p(\phi)} = \left(\Omega'(e) - \kappa\right)\left(1 - \chi(\phi)\right) \tag{79}$$

in steady state, where $\chi(\phi) = -\frac{p'(\phi)}{p(\phi)}\phi$.

We apply the envelope theorem to obtain a close-form solution for $\Omega'(e)$:

$$\Omega'(e) = \tag{80}$$
$$\frac{1}{1+rdt}\left\{g(n)dt - C'(e)dt + \Omega'(e')\left(1 - \lambda dt\right) + \left[\Omega'(e') - \kappa\right]\left(\frac{\partial\mathcal{E}}{\partial e}p(\phi) + p'(\phi)\mathcal{E}\frac{\partial\phi}{\partial e}\right)dt\right\}$$

Given that

$$\frac{\partial\mathcal{E}}{\partial e} = -(n+1)$$

and

$$\frac{\partial\phi}{\partial e} = -\frac{n+1}{\mathcal{K}},$$

we can simplify (80) as

$$\Omega'(e) =$$
$$\frac{1}{1+rdt}\left\{g(n)dt - C'(e)dt + \Omega'(e')\left(1 - \lambda dt\right) - \left[\Omega'(e') - \kappa\right](n+1)\left(p(\phi) + p'(\phi)\phi\right)dt\right\}$$

By rearranging terms in the equation above, we have

$$\Omega'(e) = \frac{g(n) - C'(e) + p(\phi)(n+1)(1-\chi(\phi))\kappa}{r + \lambda + p(\phi)(n+1)(1-\chi(\phi))}. \tag{81}$$

Thus, by replacing (81) in (79), we obtain the first-order condition for the optimal allocation of workers and fund raisers:

$$\frac{g'(n)}{p(\phi)} = (1-\chi(\phi))\left[\frac{\pi(n) - C'(e)}{r+\lambda} - \kappa\right]. \tag{82}$$

To obtain the first-order condition for $\mathcal{K}$, we derive problem (77) with respect to $\mathcal{K}$ and set the derivative to zero:

$$0 = \frac{1}{1+rdt}\left\{-\eta dt + \Omega'(e')\frac{\partial e'}{\partial\mathcal{K}} - \kappa\left(p(\phi)\frac{\partial\mathcal{E}}{\partial\mathcal{K}} + p'(\phi)\frac{\partial\phi}{\partial\mathcal{K}}\mathcal{E}\right)dt\right\}. \tag{83}$$

Given that

$$\frac{\partial e'}{\partial \mathcal{K}} = \left(p(\phi) \frac{\partial \mathcal{E}}{\partial \mathcal{K}} + p'(\phi) \frac{\partial \phi}{\partial \mathcal{K}} \mathcal{E} \right) dt$$

we can simplify (83) as

$$0 = -\eta dt + \left(\Omega'(e') - \kappa \right) \left(p(\phi) \frac{\partial \mathcal{E}}{\partial \mathcal{K}} + p'(\phi) \frac{\partial \phi}{\partial \mathcal{K}} \mathcal{E} \right) dt. \tag{84}$$

Then, notice that

$$\frac{\partial \mathcal{E}}{\partial \mathcal{K}} = 0$$

and

$$\frac{\partial \phi}{\partial \mathcal{K}} = -\frac{\mathcal{E}}{\mathcal{K}^2}$$

to get

$$0 = -\eta - \left(\Omega'(e') - \kappa \right) p'(\phi) \phi^2.$$

By rearranging terms in the equation above, we get

$$\frac{\eta}{q(\phi)} = \chi(\phi) \left(\Omega'(e) - \kappa \right). \tag{85}$$

Finally, by combining (85) with (79) and (82), we can easily obtain the first-order condition for the optimal creation of branches:

$$\frac{\eta}{q(\phi)} = \chi(\phi) \left[\frac{\pi(n) - C'(e)}{r + \lambda} - \kappa \right]. \tag{86}$$

In this simplified setting, one can easily show that the first-order conditions of the decentralized equilibrium still are (FC) and (CC). One can thus deduce if the decentralized equilibrium is efficient by comparing (86) with (CC) and (82) with (FC). The decentralized equilibrium is characterized by three inefficiencies: i) congestion externalities that require $\beta = \chi(\phi)$; ii) a hold-up problem that disappears if $\theta = 1$; iii) a scale inefficiency that disappears if $\Delta = 1$ or if agents take ρ as given when they take economic decisions.

A.12.2 Case with a fixed mass of heterogenous banks

We now consider a situation where banks differ in terms of their agency cost function. There is a mass b of banks that we index by $i \in (0, b)$. We denote by $\mathcal{M}$ the continuum

of branches M_i for all $i \in (0, b)$. The social planner problem becomes

$$\Omega(\mathcal{M}) = \max_{\{n_i\}_0^b, \{K_i\}_0^b} \frac{1}{1+rdt} \left\{ \left[\int_0^b M_i g(n_i) di - \int_0^b \left[\eta K_i + C_i(M_i) + c \right] di \right] dt + \Omega(\mathcal{M}') - \frac{\kappa p(\phi) \mathcal{E} dt}{1+rdt} \right\} \tag{87}$$

such that

$$\dot{M}_i = K_i q(\phi) - \lambda M_i,$$

with

$$\mathcal{E} = 1 - \int_0^b M_i (n_i + 1) di$$

and

$$\phi = \frac{1 - \int_0^b M_i (n_i + 1) di}{\int_0^b K_i di}.$$

Define $\mathcal{K} = \int_0^b K_i di$ and $e = \int_0^b M_i di$. By deriving (87) with respect to n_i, we obtain the following first-order condition:

$$\frac{1}{1+rdt} \left\{ M_i g'(n_i) didt + \int_0^b \frac{\partial \Omega(\mathcal{M}')}{\partial M_j'} \frac{\partial M_j'}{\partial n_i} dj - \kappa \left(p'(\phi) \frac{\partial \phi}{\partial n_i} \mathcal{E} + p(\phi) \frac{\partial \mathcal{E}}{\partial n_i} \right) dt \right\} = 0.$$

We use the notation di to refer to the (arbitrarily small) mass of a bank. Given that

$$\frac{\partial \phi}{\partial n_i} = -\frac{M_i}{\mathcal{K}} di,$$

$$\frac{\partial \mathcal{E}}{\partial n_i} = -M_i di$$

and

$$\frac{\partial M_j'}{\partial n_i} = -K_j q'(\phi) \frac{M_i}{\mathcal{K}} didt,$$

we can rewrite this first-order condition as

$$M_i g'(n_i) didt - \int_0^b \frac{\partial \Omega(\mathcal{M}')}{\partial M_j'} M_i q'(\phi) \frac{K_j}{\mathcal{K}} djdidt + \kappa M_i \left(p'(\phi)\phi + p(\phi) \right) didt = 0.$$

The equation above can then be simplified as

$$\frac{g'(n_i)}{p(\phi)} = (1 - \chi(\phi)) \left[\int_0^b \frac{\partial \Omega(\mathcal{M})}{\partial M_j} \frac{K_j}{\mathcal{K}} dj - \kappa \right] \tag{88}$$

in steady state, where $\chi(\phi) = -\frac{p'(\phi)}{p(\phi)}\phi$. Since n_i depends only upon aggregate variables, we can deduce that $n_i = n$ for all $i \in (0, b)$.

To obtain a close-form solution to the right-hand side of (88), we apply the envelope theorem to (87):

$$\frac{\partial \Omega(\mathcal{M})}{\partial M_i} di = \frac{1}{1 + rdt} \left\{ \left[g(n) - C_i'(M_i) \right] didt + \int_0^b \frac{\partial \Omega}{\partial M_j'} \frac{\partial M_j'}{\partial M_i} dj - \kappa \left(p'(\phi) \mathcal{E} \frac{\partial \phi}{\partial M_i} + p(\phi) \frac{\partial \mathcal{E}}{\partial M_i} \right) dt \right\}, \tag{89}$$

where we use the fact that $n_i = n$ for all $i \in (0, 1)$.

Notice that

$$\frac{\partial M_j'}{\partial M_i} = \begin{cases} 1 - \lambda dt - \frac{K_j}{\mathcal{K}} q'(\phi)(n+1) didt & \text{if } i = j \\ -\frac{K_j}{\mathcal{K}} q'(\phi)(n+1) didt & \text{if } i \neq j, \end{cases}$$

$$\frac{\partial \phi}{\partial M_i} = -\frac{n+1}{\mathcal{K}} di$$

and

$$\frac{\partial \mathcal{E}}{\partial M_i} = -(n+1) di.$$

Hence, (89) can be rewritten as

$$\frac{\partial \Omega(\mathcal{M})}{\partial M_i} di = \frac{1}{1 + rdt} \Bigg\{ \left[g(n) - C_i'(M_i) \right] didt - \int_0^b \frac{\partial \Omega(\mathcal{M})}{\partial M_j} \frac{K_j}{\mathcal{K}} q'(\phi)(n+1) djdidt$$
$$+ \tfrac{\partial \Omega(\mathcal{M})}{\partial M_i} (1 - \lambda dt) di + (n+1) \kappa \left(p'(\phi)\phi + p(\phi) \right) didt \Bigg\},$$

in steady state. By rearranging terms, we obtain

$$(r + \lambda) \frac{\partial \Omega(\mathcal{M})}{\partial M_i} = g(n_i) - C_i'(M_i) - (n+1) q'(\phi) \left[\left(\int_0^b \frac{\partial \Omega(\mathcal{M})}{\partial M_j} \frac{K_j}{\mathcal{K}} \right) - \kappa \right],$$

which, by use of (88) can be rewritten as

$$(r + \lambda) \frac{\partial \Omega(\mathcal{M})}{\partial M_i} = \pi(n) - C_i'(M_i). \tag{90}$$

By combining (88) and (90), we obtain

$$\frac{g'(n)}{p(\phi)} = (1 - \chi(\phi)) \left[\int_0^b \frac{\pi(n) - C_i'(M_i)}{r + \lambda} \frac{K_i}{\mathcal{K}} di - \kappa \right]. \tag{91}$$

A first-order condition for the optimal mass of branches can obtained by deriving (87) with respect to K_i and setting the derivative to zero:

$$\frac{1}{1+rdt}\left\{-\eta didt+\int_0^b \frac{\partial\Omega(\mathcal{M})}{\partial M_j'}\frac{\partial M_j'}{\partial K_i}dj-\kappa\left(p'(\phi)\mathcal{E}\frac{\partial\phi}{\partial K_i}+p(\phi)\frac{\partial\mathcal{E}}{\partial K_i}\right)dt\right\}=0.$$

Given that

$$\frac{\partial\phi}{\partial K_i}=-\frac{\phi}{\mathcal{K}}di,$$

$$\frac{\partial\mathcal{E}}{\partial K_i}=0$$

and

$$\frac{\partial Mj'}{\partial K_i}=\begin{cases}q(\phi)-\frac{K_j}{\mathcal{K}}q'(\phi)\phi di & \text{if } i=j\\ -\frac{K_j}{\mathcal{K}}q'(\phi)\phi di & \text{if } i\neq j,\end{cases}$$

we can rewrite this first-order condition as

$$\eta=\frac{\partial\Omega(\mathcal{M})}{\partial M_i'}q(\phi)-\int_0^b\frac{\partial\Omega(\mathcal{M})}{\partial M_j'}\frac{K_j}{\mathcal{K}}q'(\phi)\phi dj+\kappa p'(\phi)\phi^2,$$

or equivalently as

$$\frac{\eta}{q(\phi)}=\frac{\partial\Omega(\mathcal{M})}{\partial M_i'}-\int_0^b\frac{\partial\Omega(\mathcal{M})}{\partial M_j'}\frac{K_j}{\mathcal{K}}(1-\chi(\phi))dj-\kappa\chi(\phi)$$

or

$$\frac{\eta}{q(\phi)}=\chi(\phi)\left[\int_0^b\frac{\partial\Omega(\mathcal{M})}{\partial M_j}\frac{K_j}{\mathcal{K}}dj-\kappa\right]+\frac{\partial\Omega(\mathcal{M})}{\partial M_i}-\int_0^b\frac{\partial\Omega(\mathcal{M})}{\partial M_j}\frac{K_j}{\mathcal{K}}dj.$$

Notice that, according to the equation above, $\frac{\partial\Omega(\mathcal{M})}{\partial M_i}$ must be the same across all banks. Hence, the following property must hold:

$$\frac{\partial\Omega(\mathcal{M})}{\partial M_i}=\int_0^b\frac{\partial\Omega(\mathcal{M})}{\partial M_j}\frac{K_j}{\mathcal{K}}dj,\forall i\in(0,b),$$

implying

$$\frac{\eta}{q(\phi)}=\chi(\phi)\left[\int_0^b\frac{\partial\Omega}{\partial M_j}\frac{K_j}{\mathcal{K}}dj-\kappa.\right] \tag{92}$$

By combining (92) with (90) and the fact that $\frac{\partial\Omega(\mathcal{M})}{\partial M_i}$ must be the same across all

banks, we get

$$\frac{\eta}{q(\phi)} = \chi(\phi)\left[\frac{\pi(n)-\sigma}{r+\lambda} - \kappa\right], \tag{93}$$

where $\sigma = C_i'(M_i)$, equal across all banks.

Finally, by combining (91) with the fact that $\frac{\partial\Omega(\mathcal{M})}{\partial M_i}$ is equal across all banks yields

$$\frac{g'(n)}{p(\phi)} = (1-\chi(\phi))\left[\frac{\pi(n)-\sigma}{r+\lambda} - \kappa\right], \tag{94}$$

To conclude, conditions (93) and (94) turn out to be the same as conditions (82) and (86), which describe the social planne's solution in a context with homogenous banks. The model with bank heterogeneity is thus characterized by the same inefficiencies as the model with homogenous banks.

A.12.3 Case with bank entry and exit

Consider now the following problem to solve:

$$\Omega(\mathcal{M},b) = \max_{\{n_\varphi\}_{\varphi^*}^{\infty},\{K_\varphi\}_{\varphi^*}^{\infty},b^e,\varphi^*} \tag{95}$$
$$\frac{1}{1+rdt}\left\{\left[\int_{\varphi^*}^{\infty} M_\varphi g(n_\varphi)\frac{dF(\varphi)}{1-F(\varphi^*)} - \int_{\varphi^*}^{\infty}\left[\eta K_\varphi + C_\varphi(M_\varphi)+c\right]\frac{dF(\varphi)}{1-F(\varphi^*)}\right]bdt + \Omega(\mathcal{M}',b') - \frac{\kappa p(\phi)\mathcal{E}dt}{1+rdt} - \frac{\nu b^e dt}{1-F(\varphi^*)}\right\}$$

such that

$$\dot{M}_\varphi = K_\varphi q(\phi) - \lambda M_\varphi,$$
$$\dot{b} = b^e,$$

with

$$\mathcal{E} = 1 - b\int_{\varphi^*}^{\infty} M_\varphi(n_\varphi+1)\frac{dF(\varphi)}{1-F(\varphi^*)}$$

and

$$\phi = \frac{1 - b\int_{\varphi^*}^{\infty} M_\varphi(n_\varphi+1)\frac{dF(\varphi)}{1-F(\varphi^*)}}{b\int_{\varphi^*}^{\infty} K_\varphi\frac{dF(\varphi)}{1-F(\varphi^*)}}.$$

Define $\mathcal{K} = b\int_{\varphi^*}^{\infty} K_\varphi\frac{dF(\varphi)}{1-F(\varphi^*)}$ and $e = b\int_{\varphi^*}^{\infty} M_\varphi\frac{dF(\varphi)}{1-F(\varphi^*)}$. We denote by $\mathcal{M}$ the continuum of branches M_φ for all $\varphi \in (\varphi^*,\infty)$. By deriving (95) with respect to n_φ, we obtain the following first-order condition:

$$M_\varphi g'(n_\varphi)\frac{dF(\varphi)}{1-F(\varphi^*)}bdt + b\int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M}',b')}{\partial M'_{\hat{\varphi}}}\frac{\partial M'_{\hat{\varphi}}}{\partial n_\varphi}\frac{dF(\hat{\varphi})}{1-F(\varphi^*)} - \kappa\left(p'(\phi)\frac{\partial\phi}{\partial n_\varphi}\mathcal{E} + p(\phi)\frac{\partial\mathcal{E}}{\partial n_\varphi}\right)dt = 0.$$

We use the notation $d\varphi$ to refer to the (arbitrarily small) mass of a bank. Given that

$$\frac{\partial \phi}{\partial n_\varphi} = -b\frac{M_\varphi}{\mathcal{K}}\frac{d\varphi f(\varphi)}{1-F(\varphi^*)},$$

$$\frac{\partial \mathcal{E}}{\partial n_i} = -bM_\varphi\frac{d\varphi f(\varphi)}{1-F(\varphi^*)}$$

and

$$\frac{\partial M'_{\hat{\varphi}}}{\partial n_\varphi} = -bK_{\hat{\varphi}}q'(\phi)\frac{M_\varphi}{\mathcal{K}}\frac{d\varphi f(\varphi)}{1-F(\varphi^*)}dt,$$

we can rewrite this first-order condition as

$$M_\varphi g'(n_\varphi)\frac{dF(\varphi)}{1-F(\varphi^*)}bdt$$
$$-\int_{\varphi^*}^{\infty}\left(\frac{\partial\Omega(\mathcal{M}',b')}{\partial M_{\hat{\varphi}}}b^2q'(\phi)M_\varphi\frac{K_{\hat{\varphi}}}{\mathcal{K}}\frac{dF(\varphi)}{1-F(\varphi^*)}dt\right)\frac{dF(\hat{\varphi})}{1-F(\varphi^*)}+\kappa M_\varphi\left(p'(\phi)\phi+p(\phi)\right)\frac{dF(\varphi)}{1-F(\varphi^*)}bdt=0.$$

The equation above can then be simplified as

$$\frac{g'(n_\varphi)}{p(\phi)} = (1-\chi(\phi))\left[b\int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\hat{\varphi}}}\frac{K_{\hat{\varphi}}}{\mathcal{K}}\frac{dF(\hat{\varphi})}{1-F(\varphi^*)}-\kappa\right] \tag{96}$$

in steady state, where $\chi(\phi) = -\frac{p'(\phi)}{p(\phi)}\phi$. Since n_φ depends only upon aggregate variables, we can deduce that $n_\varphi = n$ for all $\varphi \in (\varphi^*, \infty)$.

To obtain a close-form solution to the right-hand side of (96), we apply the envelope theorem to (95):

$$\frac{\partial\Omega(\mathcal{M},b)}{\partial M_\varphi}b\frac{dF(\varphi)}{1-F(\varphi^*)} = \tag{97}$$
$$\frac{1}{1+rdt}\left\{\left[g(n)-C'_\varphi(M_\varphi)\right]\frac{dF(\varphi)}{1-F(\varphi^*)}bdt+b\int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M}',b)}{\partial M'_{\hat{\varphi}}}\frac{\partial M'_{\hat{\varphi}}}{\partial M'_\varphi}\frac{dF(\hat{\varphi})}{1-F(\varphi^*)}-\kappa\left(p'(\phi)\mathcal{E}\frac{\partial\phi}{\partial M_\varphi}+p(\phi)\frac{\partial\mathcal{E}}{\partial M_\varphi}\right)dt\right\},$$

where we use the fact that $n_\varphi = n$ for all $\varphi \in (0, b)$.

Notice that

$$\frac{\partial\phi}{\partial M_\varphi} = -b\frac{n+1}{\mathcal{K}}\frac{dF(\varphi)}{1-F(\varphi^*)},$$

$$\frac{\partial M'_{\hat{\varphi}}}{\partial M_\varphi} = \begin{cases} 1-\lambda dt-b\frac{K_{\hat{\varphi}}}{\mathcal{K}}q'(\phi)(n+1)\frac{dF(\varphi)}{1-F(\varphi^*)}dt & \text{if } \varphi=\hat{\varphi} \\ -b\frac{K_{\hat{\varphi}}}{\mathcal{K}}q'(\phi)(n+1)\frac{dF(\varphi)}{1-F(\varphi^*)}dt & \text{if } \varphi\neq\hat{\varphi}\end{cases}$$

and

$$\frac{\partial\mathcal{E}}{\partial M_\varphi} = -b(n+1)\frac{dF(\varphi)}{1-F(\varphi^*)}.$$

Hence, (97) can be rewritten as

$$\frac{\partial\Omega(\mathcal{M},b)}{\partial M_\varphi}b\frac{dF(\varphi)}{1-F(\varphi^*)}=\frac{1}{1+rdt}\Bigg\{\left[g(n)-C'_\varphi(M_\varphi)\right]\frac{dF(\varphi)}{1-F(\varphi^*)}bdt+b\frac{\partial\Omega(\mathcal{M},b)}{\partial M_\varphi}(1-\lambda dt)\frac{dF(\varphi)}{1-F(\varphi^*)}$$
$$-b^2\int_{\varphi^*}^{\infty}\left(\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\hat\varphi}}\frac{K_{\hat\varphi}}{\mathcal{K}}q'(\phi)(n+1)\frac{dF(\varphi)}{1-F(\varphi^*)}\right)\frac{dF(\hat\varphi)}{1-F(\varphi^*)}dt+(n+1)\kappa\left(p'(\phi)\phi+p(\phi)\right)\frac{dF(\varphi)}{1-F(\varphi^*)}bdt\Bigg\},$$

in steady state. By rearranging terms, we have that

$$(r+\lambda)\frac{\partial\Omega(\mathcal{M},b)}{\partial M_\varphi}b\frac{dF(\varphi)}{1-F(\varphi^*)}dt=\left[g(n)-C'_\varphi(M_\varphi)\right]\frac{dF(\varphi)}{1-F(\varphi^*)}bdt$$
$$-b^2\int_{\varphi^*}^{\infty}\left(\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\hat\varphi}}\frac{K_{\hat\varphi}}{\mathcal{K}}q'(\phi)(n+1)\frac{dF(\varphi)}{1-F(\varphi^*)}\right)\frac{dF(\hat\varphi)}{1-F(\varphi^*)}dt+(n+1)\kappa\left(p'(\phi)\phi+p(\phi)\right)\frac{dF(\varphi)}{1-F(\varphi^*)}bdt,$$

By simplifying, we obtain

$$(r+\lambda)\frac{\partial\Omega(\mathcal{M},b)}{\partial M_\varphi}=g(n)-C'_\varphi(M_\varphi)-(n+1)q'(\phi)\left[\left(b\int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M})}{\partial M_{\hat\varphi}}\frac{K_{\hat\varphi}}{\mathcal{K}}\frac{dF(\hat\varphi)}{1-F(\varphi^*)}\right)-\kappa\right],$$

which, by use of (96) can be rewritten as

$$(r+\lambda)\frac{\partial\Omega(\mathcal{M},b)}{\partial M_\varphi}=\pi(n)-C'_\varphi(M_\varphi). \tag{98}$$

By combining (96) and (98), we obtain

$$\frac{g'(n)}{p(\phi)}=(1-\chi(\phi))\left[\int_{\varphi^*}^{\infty}\frac{\pi(n)-C'_{\hat\varphi}(M_{\hat\varphi})}{r+\lambda}\frac{bK_{\hat\varphi}}{\mathcal{K}}\frac{dF(\hat\varphi)}{1-F(\varphi^*}-\kappa\right] \tag{99}$$

A first-order condition for the optimal mass of branches can obtained by deriving (95) with respect to K_φ and setting the derivative to zero:

$$\frac{1}{1+rdt}\left\{-\eta b\frac{dF(\varphi)}{1-F(\varphi^*)}dt+b\int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M}',b')}{\partial M'_{\hat\varphi}}\frac{\partial M_{\hat\varphi}}{\partial K_\varphi}\frac{dF(\hat\varphi)}{1-F(\varphi^*)}-\kappa\left(p'(\phi)\mathcal{E}\frac{\partial\phi}{\partial K_\varphi}+p(\phi)\frac{\partial\mathcal{E}}{\partial K_\varphi}\right)dt\right\}=0$$

Given that

$$\frac{\partial\phi}{\partial K_\varphi}=-\frac{\phi}{\mathcal{K}}b\frac{dF(\varphi)}{1-F(\varphi^*)},$$
$$\frac{\partial\mathcal{E}}{\partial K_\varphi}=0$$

and

$$\frac{\partial M'_{\hat{\varphi}}}{\partial K_{\varphi}} = \begin{cases} q(\phi)dt - b\frac{K_{\hat{\varphi}}}{\mathcal{K}}q'(\phi)\phi\frac{dF(\varphi)}{1-F(\varphi^*)}dt & \text{if } \varphi = \hat{\varphi} \\ -b\frac{K_{\hat{\varphi}}}{\mathcal{K}}q'(\phi)\phi\frac{dF(\varphi)}{1-F(\varphi^*)}dt & \text{if } \varphi \neq \hat{\varphi}, \end{cases}$$

we can rewrite this first-order condition as

$$\begin{aligned} &-\eta b\frac{dF(\varphi)}{1-F(\varphi^*)}dt + bq(\phi)\frac{\partial\Omega(\mathcal{M}',b')}{\partial M'_{\varphi}}\frac{dF(\varphi)}{1-F(\varphi^*)}dt \\ &-b\int_{\varphi^*}^{\infty}\left(\frac{\partial\Omega(\mathcal{M}',b')}{\partial M'_{\hat{\varphi}}}b\frac{K_{\hat{\varphi}}}{\mathcal{K}}q'(\phi)\phi\frac{dF(\varphi)}{1-F(\varphi^*)}dt\right)\frac{dF(\hat{\varphi})}{1-F(\varphi^*)} + \kappa p'(\phi)\phi^2 b\frac{dF(\varphi)}{1-F(\varphi^*)}dt = 0 \end{aligned}$$

and simplify it as

$$\eta = \frac{\partial\Omega(\mathcal{M}',b')}{\partial M'_{\varphi}}q(\phi) - \int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M}',b')}{\partial M'_{\hat{\varphi}}}\frac{bK_{\hat{\varphi}}}{\mathcal{K}}q'(\phi)\phi\frac{dF(\hat{\varphi})}{1-F(\varphi^*)} + \kappa p'(\phi)\phi^2.$$

By dividing by $q(\phi)$ on both sides, we obtain

$$\frac{\eta}{q(\phi)} = \frac{\partial\Omega(\mathcal{M}',b')}{\partial M'_{\varphi}} - \int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M}',b')}{\partial M'_{\hat{\varphi}}}\frac{bK_{\hat{\varphi}}}{\mathcal{K}}(1-\chi(\phi))\frac{dF(\hat{\varphi})}{1-F(\varphi^*)} - \kappa\chi(\phi) \tag{100}$$

or

$$\frac{\eta}{q(\phi)} = \chi(\phi)\left[\int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\hat{\varphi}}}\frac{bK_{\hat{\varphi}}}{\mathcal{K}}\frac{dF(\hat{\varphi})}{1-F(\varphi^*)} - \kappa\right] + \frac{\partial\Omega(\mathcal{M})}{\partial M_{\varphi}} - \int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\hat{\varphi}}}\frac{bK_{\hat{\varphi}}}{\mathcal{K}}\frac{dF(\hat{\varphi})}{1-F(\varphi^*)}$$

in steady state.

Notice that, according to the equation above, $\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\varphi}}$ must be the same across all banks. Hence, the following property must hold:

$$\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\varphi}} = \int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\hat{\varphi}}}\frac{bK_{\hat{\varphi}}}{\mathcal{K}}\frac{dF(\hat{\varphi})}{1-F(\varphi^*)}, \forall\varphi \geq \varphi^*,$$

implying

$$\frac{\eta}{q(\phi)} = \chi(\phi)\left[\int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\hat{\varphi}}}\frac{bK_{\hat{\varphi}}}{\mathcal{K}}\frac{dF(\hat{\varphi})}{1-F(\varphi^*)} - \kappa\right] \tag{101}$$

By combining (101) with (98) and the fact that $\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\varphi}}$ must be the same across all banks, we get

$$\frac{\eta}{q(\phi)} = \chi(\phi)\left[\frac{\pi(n)-\sigma}{r+\lambda} - \kappa\right], \tag{102}$$

where $\sigma = C'_{\varphi}(M_{\varphi})$, equal across all banks.

Similarly, by combining (99) with the fact that $\frac{\partial\Omega(\mathcal{M},b)}{\partial M_{\varphi}}$ is equal across all banks

yields

$$\frac{g'(n)}{p(\phi)} = (1-\chi(\phi))\left[\frac{\pi(n)-\sigma}{r+\lambda} - \kappa\right]. \tag{103}$$

To conclude, conditions (102) and (103) turn out to be the same as conditions (82)/(86) and (93)/(94) in the previous cases. The inefficiencies are thus the same in the general case. We now just need to show that bank entry is efficient to complete the proof.

Efficiency of bank entry. Call Γ the value of a new average bank after the ν cost has been paid, conditional on the fact that its efficiency is above the threshold φ^* (so that the planner does not discard it). As in the decentralized equilibrium, the mass of branches of an entering bank immediately jumps to its long-run level in a steady state[38]. Hence, the value of a new bank is equal to the value of an incumbent net of the initial search and firm creation costs required for M to jump to its long-run value. We call z_0 these initial costs:

$$\Gamma = \frac{1}{1+rdt}\left\{\frac{\partial\Omega(\mathcal{M},b)}{\partial b} - z_0 dt\right\}.$$

The optimal number of entering banks b^e follows

$$(1-F(\varphi^*))\Gamma = \nu.$$

Applying the envelope theorem to (95), we can write

$$\begin{aligned}\frac{\partial\Omega(\mathcal{M},b)}{\partial b} =& \frac{1}{1+rdt}\left\{\left[\int_{\varphi^*}^{\infty}(M_\varphi g(n_\varphi) - \eta K_\varphi - C_\varphi(M_\varphi) - c)\frac{dF(\varphi)}{1-F(\varphi^*)}\right]dt\right\} \\ &+ \frac{1}{1+rdt}\left\{\frac{\partial\Omega(\mathcal{M}',b')}{\partial b'} + b'\int_{\varphi^*}^{\infty}\frac{\partial\Omega(\mathcal{M}',b)}{\partial M_{\hat{\varphi}'}}\frac{\partial M_{\hat{\varphi}'}}{\partial b}\frac{dF(\hat{\varphi})}{1-F(\varphi^*)}\right\} \\ &- \frac{1}{1+rdt}\left\{\kappa\left(p'(\phi)\mathcal{E}\frac{\partial\phi}{\partial b} + p(\phi)\frac{\partial\mathcal{E}}{\partial b}\right)dt\right\},\end{aligned}$$

where

$$\frac{\partial M_{\hat{\varphi}'}}{\partial b} = K_{\hat{\varphi}}(p'(\phi)\phi + p(\phi))\frac{\partial\phi}{\partial b}dt,$$

$$\frac{\partial\mathcal{E}}{\partial b} = \frac{1}{b}(\phi\mathcal{K} - 1),$$

[38] From equation (100), one can see that M jumps immediately to its long-run value.

$$\frac{\partial \phi}{\partial b} = -\frac{1}{\mathcal{K} b}.$$

According to Euler's homogeneous function theorem, $\alpha C_\varphi(M_\varphi) = C'_\varphi(M_\varphi) M_\varphi$. Moreover, since n_φ is constant across φ as well as $C'_\varphi(M_\varphi) = \sigma$ and $\frac{\partial \Omega(\mathcal{M}', b)}{\partial M_\varphi} = \frac{\pi(n) - \sigma}{r + \lambda}$, we can rewrite this condition as

$$\begin{aligned}\frac{\partial \Omega(\mathcal{M}, b)}{\partial b} =& \frac{1}{1 + r dt} \left\{ \left[g(n) \frac{e}{b} - \eta \frac{\mathcal{K}}{b} - \frac{1}{\alpha} \sigma \frac{e}{b} - c \right] dt \right\} \\ &+ \frac{1}{1 + r dt} \left\{ \frac{\partial \Omega(\mathcal{M}', b')}{\partial b'} - \frac{1}{b} \frac{\pi(n) - \sigma}{r + \lambda} (p'(\phi)\phi + p(\phi)) dt \right\} \\ &- \frac{1}{1 + r dt} \left\{ \kappa \left(\frac{1}{b} p(\phi)(\phi \mathcal{K} - 1) - p'(\phi) \mathcal{E} \frac{1}{b \mathcal{K}} \right) dt \right\}.\end{aligned}$$

Given the immediate adjustment of M after entry, this condition reads

$$\begin{aligned}r \frac{\partial \Omega(\mathcal{M}, b)}{\partial b} =& \left[g(n) \frac{e}{b} - \eta \frac{\mathcal{K}}{b} - \frac{1}{\alpha} \sigma \frac{e}{b} - c \right] - \frac{1}{b} \frac{\pi(n) - \sigma}{r + \lambda} (p'(\phi)\phi + p(\phi)) \\ &- \frac{\kappa}{b} \left(p(\phi)(\phi \mathcal{K} - 1) - p'(\phi)\phi \right).\end{aligned}$$

in steady state.

Notice that $p'(\phi)\phi + p(\phi) = (1 - \chi(\phi)) p(\phi)$, where, as before, $\chi(\phi) = -\frac{p'(\phi)\phi}{p(\phi)}$. After some algebra, this expression can be written as

$$r \frac{\partial \Omega(\mathcal{M}, b)}{\partial b} = \left[g(n) \frac{e}{b} - \eta \frac{\mathcal{K}}{b} - \frac{1}{\alpha} \sigma \frac{e}{b} - c \right] - \frac{p(\phi)}{b} \left(\frac{\pi(n) - \sigma}{r + \lambda} - \kappa \right) (1 - \chi(\phi)) - \frac{\kappa}{b} q(\phi) \mathcal{K}.$$

Using condition (103),

$$r \frac{\partial \Omega(\mathcal{M}, b)}{\partial b} = \left[g(n) \frac{e}{b} - \eta \frac{\mathcal{K}}{b} - \frac{1}{\alpha} \sigma \frac{e}{b} - c \right] - \frac{1}{b} g'(n) - \frac{\kappa}{b} q(\phi) \mathcal{K}.$$

Using equation (98),

$$\begin{aligned}r \frac{\partial \Omega(\mathcal{M}, b)}{\partial b} =& \left((r + \lambda) \frac{\pi(n) - \sigma}{r + \lambda} + \sigma + g'(n)(n + 1) \right) \frac{e}{b} - \eta \frac{\mathcal{K}}{b} - \frac{1}{\alpha} \sigma \frac{e}{b} - c \\ &- \frac{1}{b} g'(n) - \frac{\kappa}{b} q(\phi) \mathcal{K}.\end{aligned}$$

Notice that from (102) and (103)

$$\frac{\pi(n)-\sigma}{r+\lambda}=\frac{\eta}{q(\phi)}+\frac{g'(n)}{p(\phi)}+\kappa. \tag{104}$$

Then,

$$r\frac{\partial\Omega(\mathcal{M},b)}{\partial b}=\left(\left(1-\frac{1}{\alpha}\right)\sigma\frac{e}{b}-c\right)+(r+\lambda)\frac{\pi(n)-\sigma}{r+\lambda}\frac{1}{b}\left(e-\frac{p(\phi)\mathcal{E}}{r+\lambda}\right).$$

Finally, notice that, in steady state, $\lambda e=p(\phi)\mathcal{E}$. Then, this last expression becomes

$$r\frac{\partial\Omega(\mathcal{M},b)}{\partial b}=\left(\left(1-\frac{1}{\alpha}\right)\sigma\frac{e}{b}-c\right)+r\frac{\pi(n)-\sigma}{r+\lambda}\frac{e}{b}.$$

By using the notation in (20), this can be rewritten as

$$r\frac{\partial\Omega(\mathcal{M},b)}{\partial b}=R(\tilde{\varphi})-c+r\frac{\pi(n)-\sigma}{r+\lambda}\tilde{M}.$$

The value of an average entering bank is thus

$$r\Gamma=R(\tilde{\varphi})-c. \tag{105}$$

This can be seen by calculating the initial investments z_0. Three components have to be considered among these initial costs: i) the search cost paid by the average entering bank required for M to jump to its long-run value, ii) the resulting firm creation costs and iii) the opportunity cost resulting from the fund raisers not producing for a period before they are matched to the branches. Call $\tilde{K}_0$ the mass of "empty" branches an average bank opens upon entry. Hence, z_0 can be written as

$$z_0=\eta\tilde{K}_0+\kappa q(\phi)\tilde{K}_0+\frac{g'(n)}{p(\phi)}q(\phi)\tilde{K}_0=\frac{\eta}{q(\phi)dt}\tilde{M}+\frac{\kappa}{dt}\tilde{M}+\frac{g'(n)}{p(\phi)dt}\tilde{M},$$

where the second equality in the equation above can be obtained from (9).

Hence, from (104), one can conclude that Γ can be written as in (105). The free entry condition can thus be written as in the decentralized equilibrium provided that φ^* is the same.

Efficiency of the threshold φ^*. Finally, we proceed to show that the productivity threshold φ^* in the decentralized equilibrium is efficient, provided that conditions (102) and (103) are met. The planner chooses φ^* such that the value of the marginal banks is zero in welfare terms. We denote by Υ_0 this value in flow terms:

$$r\Upsilon_0 = r\Upsilon - rz_0^* \frac{f(\varphi^*)}{1-F(\varphi^*)} bdt \tag{106}$$

In the equation above, z_0^* represents the initial search and firm creation costs required for M_{φ^*} to jump to its long-run value once a marginal bank has entered. It is multiplied by $\frac{f(\varphi^*)}{1-F(\varphi^*)}b$, i.e. the mass of marginal banks:

$$z_0^* = \frac{\eta}{q(\phi)} M_{\varphi^*} + \kappa M_{\varphi^*} + \frac{g'(n)}{p(\phi)} M_{\varphi^*}. \tag{107}$$

Υ is the value of marginal banks gross of these initial costs:

$$\begin{aligned} r\Upsilon = &\left[M_{\varphi^*} g(n) - \eta K_{\varphi^*} - C_{\varphi^*}(M_{\varphi^*}) - c\right] \frac{f(\varphi^*)}{1-F(\varphi^*)} bdt \\ &+\kappa \left[p'(\phi)\mathcal{E}\frac{\partial \phi}{\partial \varphi^*} + p(\phi)\frac{\partial \mathcal{E}}{\partial \varphi^*}\right] dt - \int_{\varphi^*}^{\infty} \frac{\partial \Omega(\mathcal{M},b)}{\partial M_\varphi} \frac{\partial M_\varphi}{\partial \varphi^*} \frac{dF(\varphi^*)}{1-F(\varphi^*)} b. \end{aligned}$$

Notice that

$$\frac{\partial \phi}{\partial \varphi^*} = \frac{f(\varphi^*)}{1-F(\varphi^*)} \frac{b}{\mathcal{K}} \left[M_{\varphi_*}(n+1) + K_{\varphi_*}\phi\right],$$

$$\frac{\partial \mathcal{E}}{\partial \varphi^*} = \frac{f(\varphi^*)}{1-F(\varphi^*)} bM_{\varphi_*}(n+1)$$

and

$$\frac{\partial M_\varphi}{\partial \varphi^*} = K_\varphi q'(\phi) \frac{d\phi}{d\varphi^*} dt.$$

Hence, given that $\frac{\partial \Omega(\mathcal{M},b)}{\partial M_\varphi} = \frac{\pi(n)-\sigma}{r+\lambda}$ for all φ, we have:

$$\begin{aligned} r\Upsilon = &\left[M_{\varphi^*} g(n) - \eta K_{\varphi^*} - C_{\varphi^*}(M_{\varphi^*}) - c\right] \frac{f(\varphi^*)}{1-F(\varphi^*)} bdt \\ &+\kappa \left[p'(\phi)\phi(M_{\varphi^*}(n+1) + K_{\varphi^*}\phi) + p(\phi)M_{\varphi^*}(n+1)\right] \frac{f(\varphi^*)}{1-F(\varphi^*)} bdt \\ &-\frac{\pi(n)-\sigma}{r+\lambda}(p'(\phi)\phi + p(\phi))(M_{\varphi^*}(n+1) + K_{\varphi^*}\phi)\frac{f(\varphi^*)}{1-F(\varphi^*)} bdt. \end{aligned}$$

After some algebra, we can write this expression as

$$\begin{aligned} r\Upsilon = &\Big(\left[M_{\varphi^*} g(n) - \eta K_{\varphi^*} - C_{\varphi^*}(M_{\varphi^*}) - c\right] \\ &-\left(\frac{\pi(n)-\sigma}{r+\lambda} - \kappa\right)(1-\chi(\phi))p(\phi)[M_{\varphi^*}(n+1) + \phi K_{\varphi^*}] - \kappa q(\phi) K_{\varphi^*}\Big) \frac{f(\varphi^*)}{1-F(\varphi^*)} bdt, \end{aligned}$$

where, as before, $\chi(\phi) = -\frac{p'(\phi)\phi}{p(\phi)}$. Using condition (103), this expression becomes

$$r\Upsilon = \left(M_{\varphi^*}(g(n) - g'(n)(n+1)) - C_{\varphi^*}(M_{\varphi^*}) - c - \left(\frac{\eta}{p(\phi)\phi} + \frac{g'(n)}{p(\phi)} + \kappa \right) p(\phi)\phi K_{\varphi^*} \right) \frac{f(\varphi^*)}{1 - F(\varphi^*)} bdt.$$

Notice that, in steady state, $p(\phi)\phi K_{\varphi^*} = \lambda M_{\varphi^*}$. Using the definition of profits $\pi(n) = g(n) - g'(n)(n+1)$ and condition (104),

$$r\Upsilon = \left(\pi(n) M_{\varphi^*} - C_{\varphi^*}(M_{\varphi^*}) - c - \frac{\pi(n) - \sigma}{r + \lambda} \lambda M_{\varphi^*} \right) \frac{f(\varphi^*)}{1 - F(\varphi^*)} bdt. \tag{108}$$

Using expressions (104), (107) and (108), we can rewrite (106) as

$$r\Upsilon_0 = \left(\pi(n) M_{\varphi^*} - C_{\varphi^*}(M_{\varphi^*}) - c - \frac{\pi(n) - \sigma}{r + \lambda} \lambda M_{\varphi^*} - \frac{\pi(n) - \sigma}{r + \lambda} r M_{\varphi^*} \right) \frac{f(\varphi^*)}{1 - F(\varphi^*)} bdt,$$

which can be written as

$$r\Upsilon_0 = \left(\sigma M_{\varphi^*} - C_{\varphi^*}(M_{\varphi^*}) - c \right) \frac{f(\varphi^*)}{1 - F(\varphi^*)} bdt.$$

Since Υ_0 has to be zero, then it has to be the case that

$$\sigma M_{\varphi^*} - C_{\varphi^*}(M_{\varphi^*}) = c$$

This condition pins down φ^* and is identical to condition (61) in the decentralized equilibrium. This completes the proof.

B Calibration

In this appendix we describe more in detail how we compute the moments used as targets for the calibration and some other numerical details.

B.1 Bank heterogeneity: the case of a Pareto distribution

We consider a Pareto distribution for φ, the cumulative distribution function of which is given by

$$F(\varphi) = 1 - \left(\frac{\varphi_0}{\varphi} \right)^{\varepsilon}.$$

and the density $f(\varphi)$ is

$$f(\varphi) = \varepsilon \varphi_0^{\varepsilon} \varphi^{-\varepsilon-1}.$$

From equation (24), we can calculate the measure of average efficiency as

$$\tilde{\varphi}^{\frac{1}{\alpha-1}} = \int_{\varphi^*}^{\infty} \varphi^{\frac{1}{\alpha-1}} \frac{\varepsilon \varphi_0^{\varepsilon} \varphi^{-\varepsilon-1}}{\left(\frac{\varphi_0}{\varphi^*}\right)^{\varepsilon}} d\varphi,$$

which, after some algebra, can be written as

$$\tilde{\varphi} = \left[\frac{-\varepsilon(\alpha-1)}{1-\varepsilon(\alpha-1)}\right]^{\alpha-1} \varphi^*. \tag{109}$$

Plugging equations (ZCP) and (109) into equation (FE), we obtain an expression for φ^*:

$$\varphi^* = \varphi_0 \left[\frac{c}{r\nu} \frac{1}{\varepsilon(\alpha-1)-1}\right]^{1/\varepsilon}. \tag{110}$$

B.2 Gini coefficient and the Lorenz curve

Here we calculate measures of concentration of branches across banks in the model. From equation (22) and the Pareto specification for the distribution F, we can obtain the cumulative distribution function of branches across banks:

$$F_M(M) = 1 - \left(\frac{M^*}{M}\right)^{(\alpha-1)\varepsilon}.$$

The Lorenz curve is obtained by inverting this distribution:

$$M(F) = \frac{M^*}{(1-F)^{\frac{1}{(\alpha-1)\varepsilon}}}$$

and the definition of a Lorenz curve:

$$L(F) = \frac{\int_0^F M(F_1) dF_1}{\int_0^1 M(F_1) dF_1}.$$

Hence,

$$L(F) = \frac{\int_0^F \frac{M^*}{(1-F_1)^{\frac{1}{(\alpha-1)\varepsilon}}} dF_1}{\int_0^1 \frac{M^*}{(1-F_1)^{\frac{1}{(\alpha-1)\varepsilon}}} dF_1}.$$

This can be simplified as follows:

$$L(F) = \frac{\int_0^F (1-F_1)^{-\frac{1}{(\alpha-1)\varepsilon}} dF_1}{\int_0^1 (1-F_1)^{-\frac{1}{(\alpha-1)\varepsilon}} dF_1}.$$

Given that $\varepsilon > \frac{1}{\alpha-1}$, the Lorenz curve thus follows

$$L(F) = 1 - (1-F)^{-\frac{1}{(\alpha-1)\varepsilon}+1}.$$

The Gini coefficient is obtained by calculating the area between the Lorenz curve and the 45 degree line:

$$Gini = \frac{1}{2(\alpha-1)\varepsilon - 1}.$$

B.3 Economies of scale

Economies of scale are measured in terms of the elasticity of the cost function $C(y)$ to changes in production y, that is:

$$\varepsilon_C = \frac{\partial C}{C} \frac{y}{\partial y}.$$

This measure is typically estimated from a cross section of banks. We would like to obtain a close-form solution for this elasticity in the model in order to calibrate some parameters of the model with the available evidence.

The flow costs $fc(\varphi)$ of a bank with efficiency parameter φ are

$$fc(\varphi) = \eta K_\varphi + C_\varphi(M_\varphi) + c + \kappa K_\varphi \phi p(\phi), \tag{111}$$

where

$$C(M_\varphi) = \frac{1}{\alpha} M_\varphi^\alpha. \tag{112}$$

Notice that

$$C(M_\varphi) = \frac{1}{\alpha} C'(M_\varphi) M_\varphi. \tag{113}$$

Equation (9) in steady state reads

$$K_\varphi = \frac{\lambda}{\phi p(\phi)} M_\varphi. \tag{114}$$

Plugging (113) and (114) in (111) yields

$$fc(\varphi) = \lambda \left[\frac{\eta}{\phi p(\phi)} + \kappa \right] M_\varphi + \frac{1}{\alpha} C'_\varphi(M_\varphi) M_\varphi + c.$$

In a simulated sample, banks have different size because they have different φ. Moreover, in real-data sample, observed banks have already chosen their desired size. We use the equilibrium size of banks in the model to account for this: by using condition (18), we can write the last expression as

$$fc(\varphi) = \left[\lambda \left(\frac{\eta}{\phi p(\phi)} + \kappa \right) + \frac{\varsigma}{\alpha \Delta} \right] M_\varphi + c. \tag{115}$$

Notice that expression (116) is linear in M_φ, where M_φ is the size chosen by a bank with efficiency φ. Log-linearizing this expression around the mean $\tilde{fc}(\varphi) = fc(\tilde{\varphi})$, we obtain

$$\hat{fc}(\varphi) = \frac{\tilde{M}_\varphi}{\tilde{fc}(\varphi)} \left[\lambda \left(\frac{\eta}{\phi p(\phi)} + \kappa \right) + \frac{\varsigma}{\alpha \Delta} \right] \hat{M}_\varphi, \tag{116}$$

where variables in hat denote log-deviations with respect to variables evaluated at $\tilde{\varphi}$. Then,

$$\varepsilon_C = \frac{\tilde{M}_\varphi}{\tilde{fc}(\varphi)} \left[\lambda \left(\frac{\eta}{\phi p(\phi)} + \kappa \right) + \frac{\varsigma}{\alpha \Delta} \right]. \tag{117}$$

Computing ε_C by means of the last expression is equivalent to computing it by running the following regression on simulated data:

$$log(fc(\varphi)) = constant + \varepsilon_C log(M_\varphi) + \varepsilon,$$

which is the standard equation usually estimated in the empirical literature.

B.4 X-efficiency

As stated in Section 5.1, the empirical literature usually estimates a bank-specific cost function of the following form

$$ln(CO)_{i,t} = \gamma_i \mathcal{F}(M_{it}, \mathbf{p}_t) + ln(x_i),$$

where CO_i are the costs of bank i, M_i is the production of bank i, $\mathbf{p}$ are prices of inputs, and $ln(x_i)$ is a multiplicative factor in the cost function that is specific to bank i. X-efficiency is measured as the ratio of this factor in the most efficiency bank to the factor in each bank, i.e.,

$$\text{X-eff}_i = exp(ln(x_{min}) - ln(x_i)).$$

In terms of our model, equation (111), together with (114), yields a flow cost $fc(M, \varphi)$ for a bank with efficiency φ equal to

$$fc(M, \varphi) = \lambda \left[\frac{\eta}{\phi p(\phi)} + \kappa \right] M + c + C_\varphi(M),$$

where we have highlighted the fact that the flow cost function depends on M.[39] Log-linearizing this last expression around $\{\tilde{M}, \tilde{\varphi}\}$, we obtain

$$\hat{fc}(M, \varphi) = \lambda \left[\frac{\eta}{\phi p(\phi)} + \kappa \right] \frac{\tilde{M}}{fc(\tilde{M}, \tilde{\varphi})} \hat{M} + \alpha \frac{C_{\tilde{\varphi}}(\tilde{M})}{fc(\tilde{M}, \tilde{\varphi})} \hat{M} - \frac{C_{\tilde{\varphi}}(\tilde{M})}{fc(\tilde{M}, \tilde{\varphi})} \hat{\varphi}, \tag{118}$$

where the variables in hats denote log-deviations from the variables evaluated at $\{\tilde{M}, \tilde{\varphi}\}$. Notice that, for an econometrician estimating (118), the last term corresponds to $ln(x_i)$. Then,

$$\begin{aligned} \text{X-eff}_\varphi &= exp(ln(x_{min}) - ln(x_\varphi)), \\ &= exp\left(\frac{C_{\tilde{\varphi}}(\tilde{M})}{fc(\tilde{M}, \tilde{\varphi})} \hat{\varphi} - \frac{C_{\tilde{\varphi}}(\tilde{M})}{fc(\tilde{M}, \tilde{\varphi})} \hat{\varphi_{max}} \right) \\ &= \left(\frac{\varphi}{\varphi_{max}} \right)^{\frac{C_{\tilde{\varphi}}(\tilde{M})}{fc(\tilde{M}, \tilde{\varphi})}}. \end{aligned}$$

The expression above is the measure of X-efficiency we can calculate from the model and compare directly with available estimates.

B.5 Scale efficiency

Scale efficiency is measured as

$$\text{S-eff}_i = \frac{AC_i^{min}}{AC_i},$$

where AC_i is the average cost of bank i, and AC_i^{min} is the minimum of the average

[39] The bank's level of production M depends on its efficiency parameter φ. We do not make this dependence explicit because efficiency is an unobservable parameter for the econometrician, who instead uses M as well as input prices to estimate the cost function fc.

cost function for bank i. In terms of our model, the average cost function for a bank with efficiency parameter φ is

$$AC(\varphi) = \frac{fc(\varphi)}{M_\varphi} = \lambda \left[\frac{\eta}{\phi p(\phi)} + \kappa \right] + \frac{c}{M_\varphi} + \frac{C_\varphi(M_\varphi)}{M_\varphi}.$$

The minimum of the average cost function for a bank with efficiency parameter φ can be found by deriving this last expression with respect to M_φ and equating the derivative to zero. This yields

$$M_\varphi^{min} = \left[\frac{\alpha}{\alpha - 1} c\varphi \right]^{1/\alpha}.$$

Then,

$$AC^{min}(\varphi) = \frac{fc(\varphi)}{M_\varphi^{min}} = \lambda \left[\frac{\eta}{\phi p(\phi)} + \kappa \right] + \frac{c}{M_\varphi^{min}} + \frac{C_\varphi(M_\varphi^{min})}{M_\varphi^{min}}$$

and scale efficiency for a bank with efficiency parameter φ is

$$\text{S-eff}_\varphi = \frac{AC_\varphi^{min}}{AC_\varphi}.$$

Average scale efficiency is given by

$$\tilde{\text{S-eff}} = \int_{\varphi^*}^{\infty} \text{S-eff}_\varphi \frac{dF(\varphi)}{1 - F(\varphi^*)}.$$

We approximate this integral through a quadrature method based on Reiter (2009), who uses a Newton-Cotes type method to obtain quadrature points and weights for a Pareto distribution.

B.6 Loan rate

To calculate the loan rate we follow Wasmer and Weil (2004). It is the interest rate R that equalizes the amount borrowed with the expected present discounted repayment on the loan:

$$\kappa = \frac{\rho}{R + \lambda}.$$

Hence,

$$R = \frac{\rho}{\kappa} - \lambda.$$

C Sensitivity: search duration for firms

Table 8: Sensitivity analysis: search duration for firms

Targeted value for search duration		Scale inefficiency included	Scale inefficiency excluded	Constrained-eff. allocation
0.333	Loan rate	0.1200	0.1078	n.a.
	Wage	1	1.0293	n.a.
	Firm size	17.0	15.5909	15.5468
	Mass of firms	1	1.0848	1.0886
	Bank size	15.03	1.5525	1.5525
	Mass of banks	1	10.5028	10.5395
	Search duration for firms	0.333	0.3286	0.1017
	Search duration for banks	0.0320	0.0324	0.1048
	Aggregate output	1	1.0240	1.0257
	Welfare	1	1.0472	1.0481
0.1	Loan rate	0.120	0.1078	n.a.
	Wage	1	1.0295	n.a.
	Firm size	17.0	15.5823	15.5689
	Mass of firms	1	1.08556	1.0866
	Bank size	15.03	1.5239	1.5239
	Mass of banks	1	10.7059	10.7172
	Search duration for firms	0.100	0.0986	0.0302
	Search duration for banks	0.0094	0.0095	0.0311
	Aggregate output	1	1.0243	1.0247
	Welfare	1	1.0473	1.0476
0.25	Loan rate	0.120	0.1078	n.a.
	Wage	1	1.0293	n.a.
	Firm size	17.0	15.5878	15.5546
	Mass of firms	1	1.0851	1.0879
	Bank size	15.03	1.5422	1.5422
	Mass of banks	1	10.5727	10.6025
	Search duration for firms	0.250	0.2464	0.0760
	Search duration for banks	0.0238	0.0241	0.0783
	Aggregate output	1	1.0241	1.0253
	Welfare	1	1.0472	1.0479

Table 9: Sensitivity analysis: search duration for firms

Targeted value for search duration		Scale inefficiency included	Scale inefficiency excluded	Constrained-eff. allocation
0.5	Loan rate	0.1200	0.1078	n.a.
	Wage*	1	1.0291	n.a.
	Firm size	17.0	15.5971	15.5313
	Mass of firms*	1	1.0844	1.0901
	Bank size	15.03	1.5730	1.5730
	Mass of banks*	1	10.3612	10.4152
	Search duration for firms	0.500	0.4929	0.1536
	Search duration for banks	0.0487	0.0494	0.1583
	Aggregate output*	1	1.0239	1.0263
	Welfare*	1	1.0470	1.0484
1	Loan rate	0.120	0.1079	n.a.
	Wage*	1	1.0287	n.a.
	Firm size	17.0	15.6154	15.4866
	Mass of firms*	1	1.0831	1.0942
	Bank size	15.03	1.6355	1.6355
	Mass of banks*	1	9.9534	10.0555
	Search duration for firms	1.000	0.9859	0.3139
	Search duration for banks	0.1016	0.1031	0.3238
	Aggregate output*	1	1.0234	1.0283
	Welfare*	1	1.0466	1.0493

Note: * variables normalized to one in the economy without the scale inefficiency.