A Macroeconomic Model with Heterogeneous Banks

Rustam Jamilov†

November 15, 2020

Job Market Paper

[Latest Version]

Abstract

I develop a non-linear, dynamic general equilibrium macroeconomic model with heterogeneous intermediaries, incomplete markets, monopolistic financial competition, bank default risk, and endogenous entry. The model nests the Gertler and Kiyotaki (2010) framework as a special case, breaks bank scale invariance, and generates a bank net worth fluctuation problem analogous to the canonical Bewley-Huggett-Aiyagari-Imrohoglu environment. Quantitatively, the model produces realistic, right-skewed cross-sectional distributions of bank assets, net worth, leverage, default risk, marginal costs, and interest margins. The impact of bank heterogeneity on the macroeconomy is disciplined with an endogenous distribution of the Marginal Propensity to Lend (MPL) - a sufficient statistic for macro elasticities with respect to a wide range of bank-level or aggregate shocks. In this environment, I study targeted credit policy experiments where monetary and fiscal authorities are allowed to affect *any individual* bank in the distribution. Macroeconomic effects depend on the instrument type and the region of the distribution which is being targeted. I find that the impact on aggregate demand is greater if equity injections are directed towards big banks. Direct lending and liquidity facilities are more effective if applied to small banks. I conclude by characterizing normative implications with size and income-dependent optimal bank regulation.

†Department of Economics, London Business School. Web: www.rustamjamilov.com. E-mail: rjamilov@london.edu. I am indebted to Hélène Rey for the invaluable guidance and support. I am grateful to the discussant, Itamar Drechsler, for very helpful comments. I thank Michele Andreolli, Fernando Broner, Nuno Coimbra, Ester Faia, Miguel Faria-e-Castro, Sigurd Galaasen, Francisco Gomes, Kyle Dempsey, Ragnar Juelsrud, Nobu Kiyotaki, Joseba Martinez, Frederic Mishkin, Tommaso Monacelli, Simon Mongey, Plamen Nenov, Tsvetelina Nenova, Elias Papaioannou, Pascal Paul, Albert Queralto, Ricardo Reis, Kenneth Rogoff, Petr Sadlacek, Andrew Scott, Vania Stavrakeva and participants of many seminars and conferences for many helpful comments and suggestions. I thank the AQR Asset Management Institute and the Wheeler Institute for Business and Development for financial support. Part of this research was conducted while I was visiting the Research Department of Norges Bank, whose hospitality I gratefully acknowledge. All errors are my own.

1 Introduction

The literature since the 2007-2008 Great Financial Crisis has a reached a consensus that the financial sector is crucial for our understanding of how the macroeconomy functions. However, the vast majority of studies over the past two decades have focused on a representative financial intermediary ("bank", henceforth) and abstracted from cross-sectional, distributional considerations. In the data, we see rich cross-sectional heterogeneity in bank size, leverage, default risk, costs of funds, interest margins, etc. Does this heterogeneity matter for the macroeconomy? How ought we think about policy design in such an environment? What are the implications of bank heterogeneity for socially optimal regulation, financial stability, and systemic risk?

This paper attempts to answer these questions within a novel quantitative, dynamic general equilibrium model with endogenous bank heterogeneity. My model builds on the canonical macro-banking framework of Gertler and Kiyotaki (2010) and Gertler and Karadi (2011) ("GK" models henceforth), which I am able to nest as a special case. I deviate from GK with two building blocs.[1] First, there is imperfect competition on the asset side of banks balance sheets. Banks intermediate differentiated capital goods which we label "financial varieties". Financial varieties represent complex financial products and services such as variable-rate loans, credit lines or derivatives - services that allow banks to remain in business while charging a markup over the cost of funds.[2] The elasticity of substitution across banks is constant across time, banks, and states of nature in the spirit of Dixit and Stiglitz (1977) and Melitz (2003).[3] Introduction of imperfect competition is motivated by recent empirical findings in Jamilov (2020b) who estimates reduced-form credit demand elasticities for the U.S. using bank branch-level data and a novel empirical approach. The average nationwide elasticity is found to be low, consistent with low competition on the asset side of banks balance sheets.

Second, banks in the model face idiosyncratic rate of return risk in the spirit of Benhabib et al. (2018). Markets are incomplete, and these idiosyncratic shocks are partially uninsurable. This assumption is motivated by the recent empirical work of Galaasen et al. (2020) who find, using administrative loan-level data from Norway, that idiosyncratic firm shocks survive portfolio aggregation and have a significant impact on bank returns and the aggregate economy. In our

[1] My model also features endogenous bank default risk. Versions of GK models with bank insolvency and illiquidity risks include Gertler et al. (2016) and Gertler et al. (2020).

[2] In contrast, non-differentiable financial goods could be proxied with conventional fixed-term commercial loans.

[3] One possible analytical micro-foundation for this setup is obtained using discrete choice theory. We follow (McFadden, 1984) closely. In our full model, there is both bank and firm heterogeneity. Heterogeneous borrowers choose which bank to borrow from. The two criteria for their discrete choice is the menu of bank-specific prices (loan rates) and amenities from which borrowers derive special utility. These amenities may proxy bank-level customer service or add-ons or transaction costs needed to reach a particular branch. It can be shown that, under certain conditions, the distribution of preferences for financial varieties maps directly to the constant elasticity of substitution across banks. See Appendix C for more details.

model, idiosyncratic risk, jointly with imperfect competition, break scale invariance and create a bank's exogenous net worth fluctuation problem analogous to the canonical Bewley-Huggett-Aiyagari-Imrohoglu environment (Bewley, 1977; Huggett, 1990; Aiyagari, 1994; Imrohoglu, 1996). Idiosyncratic shocks are a source of endogenous insolvency risk - every bank in the distribution is at risk of default. The distribution of default risk is priced into the distribution of bank-specific costs of funds by the household.[4] In equilibrium, stationary distributions of bank net worth, assets, leverage, interest margins, default risk and deposit rates emerge as part of the solution concept.

The model introduces a novel endogenous object: a distribution of the *Marginal Propensity to Lend (MPL)*. At the level of a bank, MPL measures the elasticity of the stock of assets to changes in net worth. In other words, by how much each bank in the economy increases its lending in response to an additional unit of net worth.[5] MPL heterogeneity arises endogenously in the model and summarizes the financial sector's responsiveness to a wide range of aggregate or bank-level fluctuations or policies that in any way affect the joint distribution of bank net worth and idiosyncratic returns - the two idiosyncratic states.

Integrating over all banks in the distribution, we obtain the aggregate MPL which can then be used to compute *macro elasticities*: the response of aggregate demand to bank-level or aggregate shocks. We illustrate the mechanism with the analysis of credit policies - various mechanisms that fiscal and monetary authorities can employ in order to stimulate credit supply and investment demand. Motivated by the worldwide policy response to the 2007-2008 Financial Crisis, we consider four different types of policies - equity injections, direct lending facilities, liquidity facilities, and debt guarantees. We begin with unanticipated, *aggregate* injections of equity into the banking sector. In this experiment, each intermediary in the distribution is bestowed with a unit of net worth, and we are interested in computing the impact on aggregate output. The macro elasticity depends explicitly on the MPL distribution, which we compute for our economy as well as for the GK benchmark. Bank heterogeneity is important - the macro elasticity of systematic equity injections is found to be twice as large in the baseline economy with heterogeneity than in the GK's representative-bank counterfactual.

For the remainder of the paper, we study macroeconomic effects of non-systematic credit policies where authorities can target any *individual* bank in the distribution. This is a key contribution of this paper. For each policy type, we proceed in two general steps. First, we segment the bank net worth distribution into ten quantiles (deciles). We run every experiment on *each decile*, each time

[4]There is no deposit insurance in the baseline model. Our banks, much like in the original work of Gertler and Karadi (2011) and Gertler and Kiyotaki (2010), represent risky large commercial and shadow banks who rely on uninsured short-term debt and hold imperfectly liquid assets. We do analyze the role of deposit insurance transfers when discussing credit policies.

[5]The MPL object in my model resembles Marginal Propensity to Consume heterogeneity from the Heterogeneous Agent New Keynesian (HANK) model literature (Kaplan et al., 2018).

assuming that only one quantile is impacted. Second, we compute post-policy MPL distributions and the resulting output elasticities and responses, again one per each affected quantile.

The impact on aggregate demand of targeted credit policies depends on the policy type and on the *region* of the distribution that is being affected. Policies that shift relative prices or the marginal costs of funds are generally more effective when applied to small banks. The list includes targeted lending and liquidity facilities, and deposit insurance schemes. This occurs because relaxation of marginal cost pressures benefits smaller intermediaries by more, as large banks' costs of funds are not materially higher than that of the lender of last resort. For these three credit policies, the impact on aggregate production is the highest if policy targets the lowest quantile of the bank net worth distribution. On the other hand, efficiency gains from targeted equity injections generally increase with the size of the affected intermediary. This is because, in general equilibrium, MPL is typically greater for high-net worth banks. Aggregate output can respond by 12 basis points more when equity is injected into large banks (highest net worth decile) vs small banks (smallest net worth decile), or by more than $245 billion in dollar equivalents.[6]

Quantitative policy analysis purposefully abstracts from any normative implications. I therefore also characterize socially optimal policy where the planner makes decisions on behalf of the whole banking sector. I explore optimal bank taxation schedules that target market externalities directly and can achieve constrained-efficient allocations.[7] Importantly, these policies must be size- and income-dependent because the joint distribution of bank net worth and idiosyncratic returns is a state variable in the model. In the market equilibrium, there are two main externalities. First, monopolistic competition in bank lending yields an *aggregate credit supply externality*. Banks can charge a markup over their heterogeneous costs of funds but do not internalize the impact of their decisions on aggregate returns. In the model, aggregate output is demand driven. And because firms are cash-strapped, demand is driven by financial intermediary activities. There is a general equilibrium disconnect between bank-level choices and aggregate returns. Higher credit margins push down returns on investment, which induces under-utilization of risky assets as a resource and a permanent decline in aggregate investment, output, and consumer welfare. This channel is analogous to the canonical aggregate demand externality but applied to the case of bank credit supply (Blanchard and Kiyotaki, 1987; Farhi and Werning, 2016).[8] Second, the assumption

[6] The slope of the macro elasticity schedule, conditional on an instrument type, may depend on various underlying mechanisms. I perform several robustness exercises, applied to the equity injections experiment as a case study. First, I look at the role of economies of scale in the banking sector. Efficiency gains from targeting larger banks are typically weaker with decreasing returns - the output elasticity curve is less steep. Second, I decompose the macro elasticity into contributions from idiosyncratic risk and monopolistic competition assumptions, by iteratively shutting down each channel and re-calculating the conditional elasticity in general equilibrium.

[7] A constrained-efficient allocation internalizes the impact of bank-level choices on aggregate returns *conditional* on the market structure. In other words, the planner cannot remove monopolistic competition or market incompleteness from the economy but could achieve the best possible outcome given the environmental constraints.

[8] Variants of an aggregate credit demand externality arise in various settings, such as Shleifer and Vishny (1988),

of uninsurable idiosyncratic rate of return risk in banking generates a distributional *pecuniary externality*: persistent ex-post heterogeneity in returns and the marginal values of bank net worth which cannot be equilized due to market incompleteness.[9]

Characterization of optimal policy produces three core results. First, the median optimal tax is a *subsidy* on bank returns, which undoes the under-lending externality due to imperfect financial competition. Second, the tax schedule redistributes *from* high *to* low income banks, a move that internalizes the pecuniary externality. All in all, the tax is highest for low-net worth banks with high idiosyncratic returns and lowest for high-net worth banks with low idiosyncratic returns. This pattern of redistribution mimics the shape of the MPL schedule - high-net worth banks with low returns have the largest MPL. Qualitatively, optimal policy implications are thus similar to the quantitative analysis of targeted equity injections: the social planner undoes under-lending by stimulating credit supply of those with the highest propensity to lend. Third, social planner's allocations significantly increase financial fragility and systemic risk as measured by the probability of bank default. This is the canonical financial competition-stability trade-off that arises in our general equilibrium framework endogenously (Hellman et al., 2000). In the model, bank default is not socially costly. However, with costly default net welfare gains from optimal bank regulation could be substantially lower, if not negative. Normative predictions of the model should therefore be taken with caution.

Calibration of the model is an auxiliary contribution. There are two sets of newly added parameters that must be calibrated - the elasticity of substitution across banks and the volatiltiy (and persistence) of the idiosyncratic bank return process. The elasticity of substitution is measured in a companion, stand-alone empirical paper which estimates reduced-form credit demand elasticities in the U.S. using branch-level bank interest rate data in a quasi-experimental setting (Jamilov, 2020b). The average nationwide elasticity is found to be roughly in the [1,2.4] range, low values that are consistent with low competition on the asset side of the bank balance sheet. The idiosyncratic bank return process is estimated explicitly in Galaasen et al. (2020): authors fit an AR(1) model into the sequence of firm-level idiosyncratic value-added shocks that impact bank returns and outcomes. They find that the shock is volatile (quarterly volatility in the [0.1,0.22] range) and not too persistent (quarterly autocorrelation of 0.35-0.55). The calibration approach in this paper takes a ballpark range for the aforementioned reduced-form parameters and tweaks them such that the model-generated distribution of book leverage matches U.S. bank data exactly.

Literature review. This paper relates to a long-running literature that introduces financial frictions and intermediaries into general equilibrium macroeconomic models. I build on the

Schmitt-Grohe and Uribe (2003), Korinek and Simsek (2016).

[9]Pecuniary externalities are a perennial feature in macro-finance and a small subset of salient, related papers includes Hart (1975), Greenwald and Stiglitz (1986), Geanakoplos and Polemarchakis (1985), Lorenzoni (2008), Caballero and Lorenzoni (2014), Benigno et al. (2016), Davila and Korinek (2017), Stavrakeva (2020).

canonical macro-banking setup of Gertler and Kiyotaki (2010) and Gertler and Karadi (2011), whom my model nests as special cases.[10] I use my framework to analyze the effects of targeted, bank-level credit policies.[11] This analysis is motivated by the response to the Great Financial Crisis by lenders of last resort around the world.

A few studies have emphasized the role of financial heterogeneity and rely, like this paper does, on some form of idiosyncratic risk and ex-post heterogeneity. Such papers include Corbae and D'Erasimo (2019), Bianchi and Bigio (2018), Rios Rull et al. (2020). Rios Rull et al. (2020) study countercyclical capital buffers in a partial-equilibrium setting with idiosyncratic bank default risk. In their model, aggregate returns on risky loans and bank entry are both exogenous, and there is perfect financial competition. In my model, bank-level decisions impact aggregate demand which drives the equilibrium return on capital accumulation. In addition, my model features imperfect competition and a flexible extensive margin with a time-varying number of incumbent, participating banks. Bianchi and Bigio (2018) study competitive banks' liquidity management problem in a model with idiosyncratic deposit withdrawal shocks. The risk of withdrawal generates a precautionary savings motive which incentives the banks to hold a buffer of reserves. In my model, the precautionary lending motive arises in response to uninsurable idiosyncratic rate of return shocks that generate an ex-post distribution of banks returns.

Corbae and D'Erasimo (2019) build a quantitative model that is somewhat close to ours. My paper differs from Corbae and D'Erasimo (2019) in at least three broad ways. First, they study oligopolistically competitive banks that are subject to idiosyncratic shocks on the liability side of the balance sheet.[12] In contrast, my paper develops an environment with monopolistically competitive banks and idiosyncratic rate of return shocks. Second, in our framework bank-level choices have impact on aggregate returns but there is a general equilibrium disconnect between the banks' private profit motive and aggregate consequences. This is the source of an aggregate credit supply externality that has a strong, dampening effect on investment demand. Finally, our CES credit market competition bloc is highly tractable, "portable", and can be readily enriched with additional parts and extensions. In Jamilov and Monacelli (2020), we introduce aggregate

[10]An incomplete list of other macroeconomic models with a financial sector includes Brunnermeier and Pedersen (2009), Adrian and Shin (2010, 2014), Jermann and Quadrini (2013), Brunnermeier and Sannikov (2014), He et al. (2016), Nuno and Thomas (2016), Gertler et al. (2016, 2020), etc. These frameworks largely abstract from distributional considerations in the financial sector and work with a representative financial intermediary/entrepreneur whose scale generally cannot be pinned down.

[11]Credit policy has been modelled in several representative-agent frameworks, for example (Curdia and Woodford, 2010, 2016)

[12]Other studies that look at imperfect financial competition in equilibrium setups include Stavrakeva (2019), Capelle (2019), Nguyen (2015), Christiano and Ikeda (2013), Davydiuk (2020), Rios Rull et al. (2020) and Martinez-Miera and Repullo (2010). Corbae and Levine (2018) review the state of the literature on financial competition in their 2018 Jackson Hole Symposium address. Their work stresses the theoretical interactions between competition, financial fragility, and monetary policy.

uncertainty into a variant of my framework and study the dynamics of the distribution of bank size and risk over the business cycle.[13]

My model is also related to the important work of Coimbra and Rey (2019) who develop a general equilibrium model with heterogeneous intermediaries and endogenous financial stability. Their model features, like ours, dynamic intensive and extensive margins of bank risk-taking. Our approach differs from theirs in two substantial ways. First, in my model market incompleteness and uninsured idiosyncratic return risk achieve persistent *ex-post* heterogeneity of bank returns. In Coimbra and Rey (2019), heterogeneity is achieved through an *ex-ante* distribution of bank value-at-risk constraints. Second, our model departs from the assumption of perfect competition in bank lending. This channel delivers rich ex-post heterogeneity in marginal costs and relative prices.[14]

The rest of the paper is structured as follows. Section 2 lays out the model. Section 3 discusses how we take the model to the data. Section 4 studies quantitative credit policy experiments. Section 5 analyzes normative implications of the model and computes the welfare costs of local credit market power. Finally, Section 6 concludes.

2 Model

In this section, I lay out the model, discuss its key building blocks, and analyze equilibrium properties.

2.1 Overview

Time is discrete and infinite. The economy consists of a representative household, a continuum of financial intermediaries, a representative final goods producer, and a capital goods producer. The final good is produced by a representative, perfectly competitive firm that uses aggregate capital and labor as inputs. Labor is supplied by the household inelastically. Aggregate capital is produced by a representative capital goods producer which is cash-strapped at the beginning of the period. This creates a role for banks. Banks intermediate all funds in the economy and have both marginal and fixed costs of operation. Banks start the period by buying claims on the aggregate capital stock, which the capital goods firm uses to assemble the aggregate capital bundle. Competition is monopolistic and capital goods are imperfect substitutes across banks. We label these goods "financial varieties", which proxies complex financial services such as variable-rate loans, credit

[13] A growing literature also looks at imperfect competition in the market for bank deposits and liabilities in general, a channel that we abstract from in this paper (Drechsler et al., 2017; Egan et al., 2017)

[14] Other papers that develop models with ex-ante financial heterogeneity include Korinek and Nowak (2016), Boissay et al. (2016), Begenau and Landvoigt (2018), Begenau et al. (2020), Goldstein et al. (2020).

lines, research products, derivatives. etc. The elasticity of substitution across banks is constant across time and states of nature. In this market structure, a financial intermediary can charge a markup over the bank-specific marginal cost. We provide a possible microfoundation for this market structure in Section C. Upon assembly completion, the aggregate capital stock gets transferred to banks who then rent it to the final goods firm at competitive rates. Capital depreciates fully every period after it is used in production.

Banks accumulate net worth and can borrow from households in the form of short-term debt subject to a moral hazard friction that puts an endogenous limit on the leverage ratio. Portfolio return of the banker consists of a systematic component (realized return on aggregate capital) and a persistent, idiosyncratic rate of return process. Trading markets are incomplete and these shocks are partially uninsurable. Although all banks are ex-ante homogenous, idiosyncratic risk generates ex-post heterogeneity in returns which then feeds into balance sheet heterogeneity. Negative realizations of idiosyncratic returns can drive any bank into insolvency. The household internalizes bank default risk and prices it competitively into the distribution of deposit rates. There is no deposit insurance in the baseline economy, and we add it at a later stage. Bank entry and the mass of active banks are both endogenous. Entering financiers pay a fixed startup cost and obtain a one-time idiosyncratic return draw together with some startup equity. Having observed the draw, financiers decide whether to operate or to immediately exit. Operating financiers become financial varieties and add to the mass of incumbents. The mass of entering financiers grows until bank profits (in expectation) remain above the startup costs.

Uninsurable idiosyncratic return risk and imperfect competition break down scale invariance, thereby generating an endogenous cross-section of bank assets. Equilibrium is also associated with ergodic distributions of bank net worth, leverage, default risk, loan margins, and deposit rates.

2.2 Environment

Final Good Production The final good is produced from aggregate capital and labor using a Cobb-Douglas technology. Aggregate capital is borrowed from the banks in return for realized returns on capital defined below:

$$Y_t = AK_t^{\alpha} L_t^{1-\alpha} \tag{1}$$

Wages and capital returns are competitive and follow directly from the production function and firms' optimization. There is no aggregate uncertainty in the model and capital depreciates fully every period. Systematic returns on aggregate capital R_t^k are thus:

$$R_{t+1}^k = \frac{A\alpha K_{t+1}^{\alpha-1}}{P_t} \tag{2}$$

Capital Goods Producer Suppose that capital goods are intermediated via sophisticated financial services such as variable-rate loans, credit lines, financial derivatives, equity or fixed income research products, and any other service where a bank can exercise some form of discretion and skill. A representative, perfectly competitive capital producing firm begins the period by issuing claims to the banks in return for the aggregate capital bundle. It makes zero profit every period.

There exists a time-varying mass H_t of banks, indexed by (j). Each bank supplies a unique amenity or variety that the borrower derives special utility from. The elasticity of substitution across banks ($\theta > 1$) is constant. Differentiated capital goods are assembled by the capital goods firm using a Dixit-Stiglitz aggregator from a mass H_t of available financial varieties k(j) where $j \in [0, H_t]$:

$$K_t = \left[\int_0^{H_t} k_t(j)^{\frac{\theta-1}{\theta}} dj\right]^{\frac{\theta}{\theta-1}} \tag{3}$$

The corresponding maximization problem of the capital goods firm is:

$$\max_{k_t(j)} \left[P_t K_t - \int_0^{H_t} p_t(j) k_t(j) dj\right]$$

subject to technology 3. This yields a downward-sloping demand function for bank funds:

$$k_t(j) = \left(\frac{p_t(j)}{P_t}\right)^{-\theta} K_t \tag{4}$$

Where P_t is the true relative price index of differentiated goods:

$$P_t := \left[\int_0^{H_t} p(j)^{1-\theta} dj\right]^{\frac{1}{1-\theta}} \tag{5}$$

2.3 Microfounding Monopolistic Competition in Bank Lending

I now briefly summarize how one can theoretically underpin the CES credit supply system above. My approach follows discrete choice theory where each borrower chooses both the size of the loan and the "type" of the bank to borrow from (McFadden, 1984). Appendix C provides a detailed analytical guide. The approach generalizes the case of a representative capital goods producer to a large number of borrowing firms that are heterogeneous in their preferences for banks/branches. In other words, there are firm-bank fixed-effect shocks. These shocks are cross-sectionally correlated and the degree of correlation will map intuitively to the constant elasticity θ.

Credit market power at the level of a bank can now be viewed as being isomorphic to consumer (firms, in this case) preferences for branch services and amenities or disutility towards transportation and travel that is required to reach the preferred/ideal bank or branch. Even if a particular bank offers higher loan rates (prices on claims), it can still remain in business if borrower-bank-specific preference shocks are sufficiently dispersed. The problem of heterogeneous firms is static. In our dynamic setting, as long as the distribution of preferences is not dynamic or aggregate state-dependent, the identical problem would yield the same solution every period. We therefore proceed with working with this representative-firm approximation of the more sophisticated heterogeneous-firms environment that is understood to be operating in the background.

2.4 Banks

The credit demand system in (3)-(5) is taken as given by every bank. Intermediaries start the period with initial net worth $n \in \mathbf{N} \subset \mathbf{R}_{+}$ and must choose the stock of assets $k(j)$, deposits $d(j)$, and price of varieties $p(j)$ while respecting the balance sheet constraint:

$$d_t(j) + n_t(j) = p_t(j)k_t(j) \tag{6}$$

Every bank faces fixed costs of operations FC and non-interest expenses $\frac{1}{\zeta_1}k_t(j)^{\zeta_2}$ where parameter ζ_2 can help govern the degree of increasing vs decreasing returns to scale. When choosing the size of the balance sheet, banks can borrow deposits $d(j)$ from the household, subject to the bank-specific interest rate $\bar{R}_t(j)$.

At the end of each period, every bank earns realized returns on claims on the final goods firm. Each bank earns a portfolio return $R^T_t(j)$ that comprises the return on aggregate capital R^k_t, which is common to all j, and an idiosyncratic component $\xi_t(j)$ which is specific to each bank:

$$R^T_t(j) = \kappa\xi_t(j) + (1-\kappa)R^k_t \tag{7}$$

Where $0 < \kappa < 1$ is a parameter that governs the ability to hedge idiosyncratic shocks. The idiosyncratic return, $\xi \in \Xi$, follows an AR(1) process:

$$\xi_t(j) = (1-\rho_\xi)\mu_\xi + \rho_\xi\xi_{t-1}(j) + \sigma_\xi\epsilon_t(j) \tag{8}$$

Analogously, let the finite state Markov representation of (8) be $\mathbf{G}(\xi_{t+1}, \xi_t)$. The law of motion of bank net worth is:

$$n_{t+1}(j) = R^T_{t+1}(j)p_t(j)k_t(j) - \bar{R}_t(j)d_t(j) - \frac{1}{\zeta_1}k_t(j)^{\zeta_2} - FC \tag{9}$$

Following Gertler and Karadi (2011) and Gertler and Kiyotaki (2010), there is a moral hazard problem in the deposit market. The bank has an incentive to divert franchise assets with the ability to divert no more than a fraction λ of the total value of revenues $p(j)k(j)$. If deciding to divert, the banker always escapes but the franchise enters bankruptcy the following period. The banker is indifferent between operating honestly and diverting when whatever he is able to finance exactly equals the value of the franchise. This yields the following incentive constraint that puts a limit on bank leverage.

$$\lambda p_t(j)k_t(j) \leq V_t(j) \tag{10}$$

where $V_t(j)$ is the franchise value of the intermediary, to be defined below. Each bank can default with its own probability $\upsilon(j)$. Default risk is due to fundamental insolvency, i.e. when net worth at normal market prices is non-positive[15]:

$$\upsilon_t(j) = \Pr\Big(n_{t+1}(j) \leq 0\Big) \tag{11}$$

Conditional on insolvency, the household recovers a fraction of promised payments $x_t(j)$, an object that we define later. Because at normal market prices the recovery rate $x_t(j)$ is increasing in bank size, insolvency risk is concentrated in the *left* tail of the stationary bank size distribution.

Let $\mu(n, \xi)$ be a probability measure, defined on the Borel algebra B that is generated by open subsets of the product space $\mathbf{B} = \mathbf{N} \times \Xi$, corresponding to the distribution of incumbent banks with net worth n and idiosyncratic return realizations ξ. The law of motion for the distribution is:

$$\mu_{t+1}(n_{t+1}, \xi_{t+1}) = \Gamma(\mu_t, M_{t+1}) \tag{12}$$

with M_{t+1} the total mass of entrants in period $t + 1$. M_{t+1} is determined in equilibrium by the optimal entry problem that we discuss later. We define Γ in detail below.

Dynamic Problem of the Incumbent Banker The following summarizes the dynamic problem of the incumbent. We adopt recursive notation because the solution does not depend on a specific bank j but on the relevant state variables only. Define $\mathbf{s} = \{n, \xi\}$ as the bank's idiosyncratic state vector. There is no aggregate risk. The bank maximizes its franchise value which is defined as the discounted stream of future flows of net worth. With an exogenous probability σ the incumbent may exit involuntarily, in which case his net worth gets transferred lump sum to the household. The banker discounts the future by adopting and augmenting the household's stochastic discount factor Λ, which is determined in equilibrium and defined later. Each banker takes as given aggregate quantities $\{K\}$, prices $\{P, R^T(\mathbf{s})\}$, the cross-sectional distribution and number of entrants $\{\mu, M\}$,

[15]Liquidity crises and bank run risk are explored in a related paper that incorporates targeted, idiosyncratic bank runs into this framework Jamilov (2020a).

bank-specific deposit rates $\bar{R}(\mathbf{s})$, and the law of motion of the distribution Γ. Each bank solves:

$$V(\mathbf{s}) = \max_{\{k,p,d\}} \left\{ \mathbb{E}_{\mathbf{s'|s}} \left[\max\left(\Lambda' \left((1-\sigma) n' + \sigma V(\mathbf{s'}) \right), 0 \right) \right] \right\} \tag{13}$$

s.t. conditions 1-12.

Where the max operator inside captures limited liability of banks. We can simplify the problem above considerably by reformulating it into a one-argument problem. Each bank now chooses the leverage ratio $\phi = \frac{pk}{n}$ by maximizing:

$$\max_{\phi} \left[\mu_a \phi + \nu_a \right] \tag{14}$$

subject to the same constraints as before and where $\mu_a = (1-\upsilon)\tilde{\Lambda}' \left[R'^{T} - \bar{R} \right]$ is the excess return on risky investments, $\nu_a = (1-\upsilon)\tilde{\Lambda}' \left[\bar{R} - \frac{\frac{1}{\zeta_1} k_t(j)^{\zeta_2} + FC}{n} \right]$ is the cost of liabilities. In both instances, $\tilde{\Lambda}' = \Lambda \left(1 - \sigma + \sigma V(\mathbf{s'}) \right)$ is an augmented stochastic discount factor.

Bank Leverage When internalizing the credit demand system and substituting out relative prices, the leverage ratio ϕ can be shown to be:

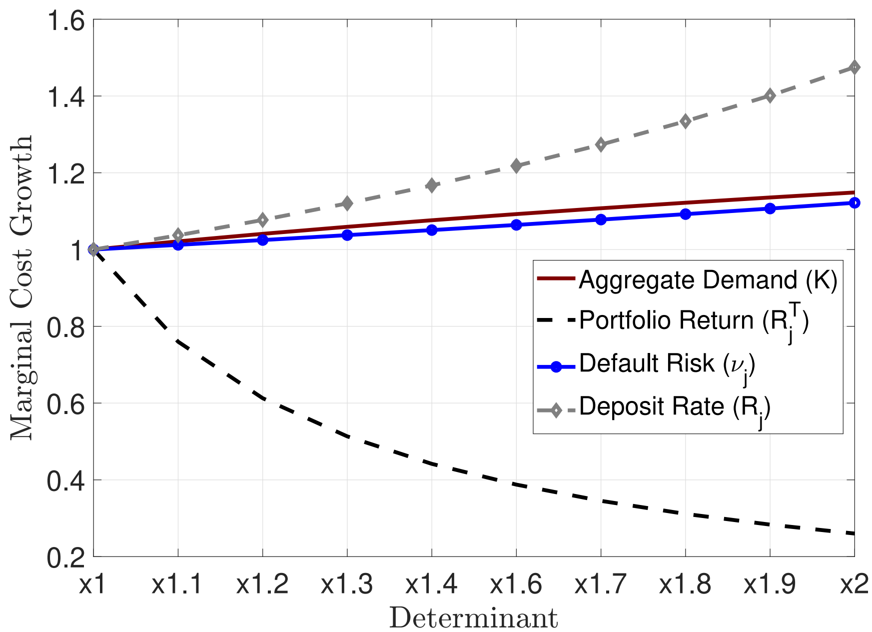
$$\phi = \frac{pk}{n} = \left(\frac{k}{K} \right)^{-\frac{1}{\theta}} \frac{Pk}{n} \tag{15}$$

The slope of the leverage curve depends directly on the degree of credit market competition, θ. In the special case of $\theta \to \infty$ and $\kappa = 0$ (as well as no non-interest and fixed expenses), the environment reduces to a homogenous good case with market leverage reverting to $\phi = \frac{Pk}{n}$, which is the definition of leverage in Gertler and Kiyotaki (2010) and Gertler and Karadi (2011).

2.5 Marginal Cost Heterogeneity

We pause the presentation of the model and zoom in on the problem of the incumbent bank. We revert back to the (j) representation temporarily to emphasize heterogeneity. It can be shown that the partial-equilibrium solution to the banks problem yields the following relative price rule:

Figure 1: **Marginal Cost Heterogeneity**



Notes: This figure plots the relationship between bank-level marginal costs and its four key components as in Equation 2.5

$$\frac{p(j)}{P} = \left[\underbrace{\frac{\theta}{\theta - 1}}_{\text{Markup}} \underbrace{\frac{1}{\Big(1 - \nu(j)\Big) R^{T}(j) - \bar{R}(j)} \frac{1}{P} K^{\zeta_2 - 1}}_{\text{Marginal Cost}} \right]^{\frac{1}{1+\theta(\zeta_2-1)}}$$

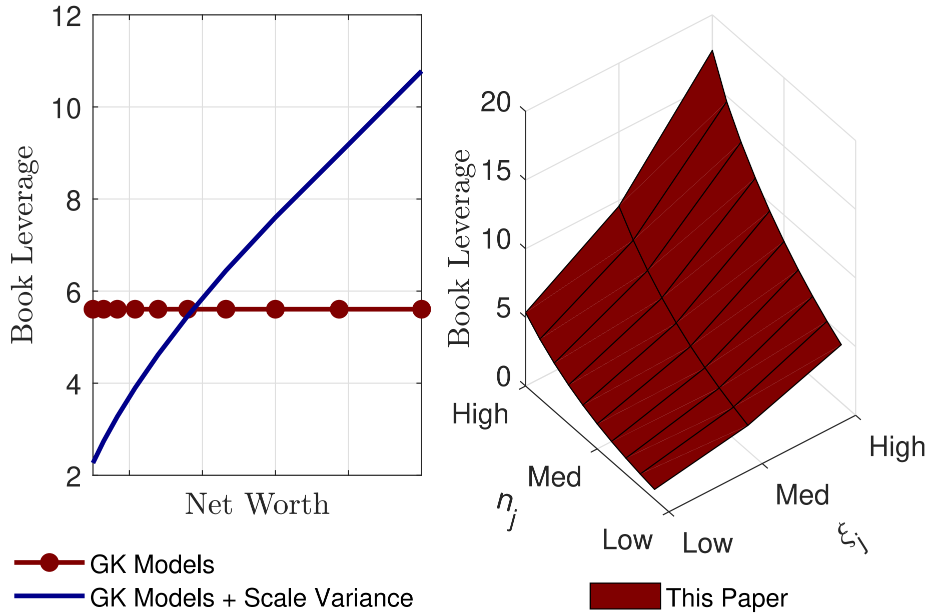
Each bank in the distribution charges a constant markup over the cost of funds. The markup is *homogenous* across banks and is equal to $\frac{\theta}{\theta-1}$, a standard formula that arises from CES monopolistic competition models.

All cross-sectional heterogeneity in the model runs through marginal costs. The marginal cost is a complex non-linear function of four key objects: the probability of default $\nu(j)$, idiosyncratic portfolio return $R^T(j)$, interest rate on deposits $\bar{R}(j)$, and the scale effect in K. Note that dependency on aggregate demand K is only possible when non-interest expenses are not linear with respect to k(j).[16]

The effect of each of the four determinants on the marginal cost is summarized in Figure 1. First, we observe that bank-level marginal costs are increasing in aggregate demand. Greater demand

[16]In the equation, we set $\zeta_1 = 1$ for simplicity and without loss of generality.

Figure 2: **Nesting Gertler and Kiyotaki (2010) and Gertler and Karadi (2011)**



Notes: This figure demonstrates the step-by-step nesting of Gertler and Kiyotaki (2010) and Gertler and Karadi (2011) models.

for bank finances puts upward pressure on the cost of funds. Second, marginal costs increase in both default risk and interest rates on deposits. Because there is no deposit insurance in the baseline economy, the two are intricately linked and have the same effect on the total marginal cost. In anticipation of equilibrium, smaller banks (low net worth) will have greater probability of default, which feeds into higher deposit rates and high marginal costs. Finally, the marginal cost is decreasing in the portfolio return $\mathrm{R}^{\mathrm{T}}(\mathrm{j})$ which acts as an exogenous "productivity" shifter.

All in all, marginal costs are lowest for banks with high net worth and idiosycnratic returns. This relationship is essential for the quantification of the model and the future discussion of aggregate elasticities. Responses to the *same* exogenous bank-level or aggregate shocks will differ in equilibrium because of this cross-sectional marginal cost heterogeneity.

2.6 Nesting GK Models

In this section we demonstrate how our incumbent banker's problem nests the Gertler and Kiyotaki (2010) and Gertler and Karadi (2011) environment as a special case. We visualize the departure graphically on figure 2. We analyze the optimal choice of bank book leverage $\frac{\mathrm{k}}{\mathrm{n}}$ in three different situations. First, we start with the GK environment with perfect competition ($\theta \rightarrow \infty$), complete markets ($\sigma_\xi$=0 and κ=0), and no non-interest expenses. As can be seen from the figure, linearity and complete markets deem the leverage ratio one-dimensional and independent of the state of initial net worth. The banking sector is comprised of a representative homogenous intermediary.

Second, the upward-sloped line on the left panel of Figure 2 plots optimal leverage for an extension of GK that allows for imperfect competition (finite $\theta > 1$) and scale variance ($\zeta_2 \neq 1$). Notice how book leverage is increasing in net worth. This relationship is driven by the fact that marginal costs decrease with the size of the intermediary: large banks face lower marginal costs, charge lower relative prices p(j), and extend more loans in book values. This relationship is consistent with the empirical evidence in Coimbra and Rey (2019) and Jamilov and Monacelli (2020). Third and finally, in the right panel of the Figure, we relax the assumption of market incompleteness. This step introduces ex-post heterogeneity in returns. Moreover, because we maintain scale variance and imperfect competition, the optimal leverage ratio now depends on two states: ξ(j) and n(j): larger, more profitable intermediaries are more risky in book leverage terms.

An important feature of the GK class of models is linearity with respect to net worth. This assumption allows the model to be aggregated explicitly. We can formalize our departure from GK by formally proving that the value function of the bank in our model is *not* linear in net worth. In this case, one must track the two-dimensional state of net worth and idiosyncratic risk in addition to aggregate factors such as the aggregate capital stock, because bank-specific characteristics matter for the choice of {k(j), p(j), d(j)}. This result is in direct contrast to the standard proofs in Gertler and Kiyotaki (2010) and Bocola (2016), among many others. Proposition 1 formalizes the intuition.

Proposition 1 (Bank Scale Variance)**.** *The solution to the incumbent banker's problem for each j, conditional on initial net worth* n(j) *and idiosyncratic return* ξ(j) *is*

$$\mathrm{V}\Big(\mathrm{n(j)}, \xi\mathrm{(j)}\Big) = \vartheta\Big(\mathrm{n(j)}, \xi\mathrm{(j)}\Big)\mathrm{n(j)}$$

where the marginal value of net worth is:

$$\vartheta\Big(\mathrm{n(j)}, \xi\mathrm{(j)}\Big) = \frac{\Big(1 - \upsilon\mathrm{(j)}\Big)\mathbb{E}\left(\Lambda'\Big[1 - \sigma + \sigma\vartheta\Big(\mathrm{n'(j)}, \xi'\mathrm{(j)}\Big)\Big]\left(\bar{\mathrm{R}}\mathrm{(j)} - \frac{\frac{1}{\zeta_1}\mathrm{k(j)}^{\zeta_2}}{\mathrm{n(j)}}\right)\right)}{1 - \varphi\Big(\mathrm{n(j)}, \xi\mathrm{(j)}\Big)}$$

and the multiplier on the moral hazard leverage constraint is

$$\varphi\Big(\mathrm{n(j)},\xi(\mathrm{j})\Big) = \max\left[1 - \frac{\Big(1-\upsilon(\mathrm{j})\Big)\mathbb{E}\left(\Lambda'\Big[1-\sigma+\sigma\vartheta\Big(\mathrm{n}'(\mathrm{j}),\xi'(\mathrm{j})\Big)\Big]\left(\bar{\mathrm{R}}(\mathrm{j}) - \frac{\frac{1}{\zeta_1}\mathrm{k(j)}^{\zeta_2}}{\mathrm{n(j)}}\right)\right)\mathrm{n(j)}}{\lambda \mathrm{k(j)}^{\frac{\theta-1}{\theta}}\mathrm{K}^{\frac{1}{\theta}}\mathrm{P}}, 0\right]$$

Proof: Appendix A.

The above proposition shows that the value function is not linear in net worth. This can be seen from the explicit dependency of the marginal value of net worth ϑ on both ξ(j) and n(j). The former is guaranteed by $\kappa > 0$ and $\sigma_\xi > 0$, i.e. idiosyncratic return risk and market incompleteness. The latter is driven by non-linearity of non-interest expenses in assets under management ($\zeta_2 \neq 1$). As a result, explicit aggregation in the banking sector is not possible as the homogeneity assumption is not satisfied. Financial intermediaries are ex-post heterogeneous in terms of returns, which feeds into all other balance sheet and income statement characteristics.

2.7 Entry and Exit

There is infinite mass of aspiring financiers who specialize in banking services. Before entry, every financier pays a fixed equity issuance cost e in units of capital. Having paid the sunk cost, the financier receives an idiosyncratic return profitability draw $\xi_0 \in \Xi$ from the ergodic distribution $\mathrm{G}_0(\xi)$ that is implied by the ξ process. The financier is also bestowed with an initial level of net worth n_0 which is a constant fraction of the aggregate stock of net worth N. Afterwards, the financier decides whether to operate or to immediately exit. Conditional on its state $\{\mathrm{n}_0, \xi_0\}$, the financier operates if and only if its expected discounted franchise value exceeds e. The value function of the entering financier is therefore:

$$\mathrm{V}^{\mathrm{e}}(\mathrm{n}_0,\xi_0) \equiv \max\left[\mathrm{V}(\mathrm{n}_0,\xi_0) - e, 0\right] \tag{16}$$

If the financier decides to state, he becomes an incumbent j, which adds to the mass H_t. Free entry drives the future expected excess value of the entering intermediaries, net of startup costs, to 0. A financier's incentive to enter is driven by the desire to earn economic profit. Entry keeps occurring until expected bank profits are equalized with the cost of financial variety origination. In equilibrium, either V^e is equal to 0, the number of entrants is 0, or both.

The incumbent intermediary is subject to two sources of exit risk: involuntary homogenous exit rate σ and the endogenous probability of default ν(j), which is bank-specific. If a bank exits, the exiting bank's market will never be taken over by any of the incumbents.

2.8 Cross-Sectional Banking Distribution

Define E_t as the mass of banks that exit the economy due to endogenous default. Recall that (1-σ) is the fraction of banks that draw an exogenous exit shock. They are replaced by the entrants M. The distribution of incumbent banks evolves according to:

$$\mu'(n', \xi') = \underbrace{\left(\sigma - E'\right) \sum_{\xi} G(\xi', \xi) \int \mathbb{1}_{\{(n,\xi)|K(n,\xi)\in \mathbf{B}\}} \mu(dn, d\xi)}_{\text{Surviving Incumbents}} + \underbrace{M' \int \mathbb{1}_{\{(n_0,\xi)|K(n,\xi)\in \mathbf{B}\}} G_0(\xi)}_{\text{New Entrants}} \tag{17}$$

Where $\mathbb{1}$ is the indicator function that takes the value of unity when the argument {.} is true and zero otherwise. Recall that $G_0(.)$ is the CDF of ξ for entering banks and $G(x', x)$ is the Markov chain for ξ of the incumbents.

2.9 Households

For simplicity, assume inelastic labor supply normalized to 1. The representative household is tasked with choosing the supply of deposits to each local credit market $b_t(j)$ and consumption C_t, subject to the standard balance sheet constraint.

$$\max_{C_t, b_t(j)} \left[\mathbb{E}_t \sum_{t=1}^{\infty} \beta^t u\left(C_t\right)\right] \quad \text{s.t.}$$

$$C_t + \int_0^{H_t} b_t(j)dj \leq W_t + \int_0^{H_t} \bar{R}_t(j) b_{t-1}(j)dj + \int_0^{H_t} \pi_t(j)dj$$

Where π are any profits (net of startup equity injections) from bank ownership which get redistributed back to the household lump sum. First order conditions for deposits yield the following equation:

$$\bar{R}_t(j) = \frac{1 - \upsilon_t(j) x_t(j) \mathbb{E}\left(R^T_{t+1}(j)\Lambda_{t+1}\right)}{\left(1 - \upsilon_t(j)\right)\mathbb{E}\left(\Lambda_{t+1}\right)} \tag{18}$$

Where $\Lambda_{t+1} = \beta \frac{u'(c_{t+1})}{u'(c_t)}$ is the stochastic discount factor. Deposits are risky because of possible bank default and absence of deposit insurance schemes. The consumer acknowledges default risk and demands a menu of deposit rates, which depend on the deposit recovery rate $x_t(j)$:

$$x_t(j) = \min\left[\frac{\phi_t(j)}{\phi_t(j) - 1}, 1\right]$$

With ϕ the market leverage ratio, defined as before.

2.10 Stationary Industry Equilibrium

Credit market clearing requires:

$$\underbrace{\mathrm{K}}_{\text{Aggregate Supply}} = \underbrace{\int_{\mathbf{B}} \Big(\mathrm{k}(\mathrm{n}, \xi)\Big)\mu(\mathrm{dn}, \mathrm{d}\xi)}_{\text{Incumbent Demand}} + \underbrace{\mathrm{M}\int_{\mathbf{B}} \Big(\mathrm{k}(\mathrm{n}_0, \xi_0)\Big)\mathrm{dG}(\xi_0)}_{\text{Entrants Demand}} + \underbrace{\mathrm{M}e}_{\text{Entry Cost}} \tag{19}$$

Where the first term on the right hand side is demand of incumbents, the second term is demand of the entrants, and the final term is the entry cost paid by the entrants. Similarly, clearing the deposit market requires:

$$\underbrace{\int_0^{\mathrm{H}} \mathrm{b(j)dj}}_{\text{Household Supply}} = \underbrace{\int_{\mathbf{B}} \Big(\mathrm{d}(\mathrm{n}, \xi)\Big)\mu(\mathrm{dn}, \mathrm{d}\xi)}_{\text{Banks Demand}} \tag{20}$$

The goods market requires the final good to be used either for household consumption or firm investment. The latter includes investment demand that is intermediated both by the incumbent and the entering bankers:

$$\mathrm{Y} = \mathrm{C} + \mathrm{I}$$

We consider equilibria without aggregate uncertainty such that all aggregate quantities, prices, and measures are time-invariant. A *Stationary Industry Equilibrium* is defined as a set of functions that include the value function of the banker $\mathrm{V}(\mathbf{s})$, optimal policies for bank capital investment $\mathrm{k}(\mathbf{s})$ and deposit demand $\mathrm{d}(\mathbf{s})$, household consumption $\mathrm{c}(\mathrm{b}_{-1})$ and deposit supply $\mathrm{b}\Big(\mathrm{b}_{-1}(\mathrm{j})\Big)$, the mass of new bank entrants M, competitive wage W and capital R^{k} pricing functions, the aggregate price of capital P, a marginal utility process Λ, and the menu of market-clearing deposit rates $\bar{\mathrm{R}}(\mathbf{s})$ such that:

1. The household's choices $\{\mathrm{C}, \mathrm{b(j)}\}$ are optimal conditional on $\{\mathrm{W}, \bar{\mathrm{R}}(\mathrm{j})\}$
2. The banks choices $\{\mathrm{k}, \mathrm{p}, \mathrm{d}\}$ are optimal conditional on $\{\Lambda, \mathrm{K}, \mathrm{P}, \bar{\mathrm{R}}(\mathrm{j}), \mu\}$
3. Returns on factors of production are: $\mathrm{R}^{\mathrm{k}} = \frac{\alpha \mathrm{AK}^{\alpha-1}}{\mathrm{P}}$, $\mathrm{W} = (1-\alpha)\mathrm{AK}^{\alpha}$
4. $\{\mathrm{K}, \mathrm{D}, \mathrm{N}\}$ are consistent with the cross-sectional distribution and the monopolistic credit demand system in (3)-(5)
5. The free-entry condition (16) is satisfied and is consistent with individual choices

6. The credit market clears as in (19). The deposit market clears as in (20)

7. The cross-sectional distribution evolves according to (17) and is consistent with bank-level demand functions

Model Solution The numerical algorithm that I use to solve the model is described in Appendix D. Sections B.1 and B.2 illustrate policy functions and bivariate cross-sectional distributions that are consistent with the stationary industry equilibrium.

2.11 Symmetric Equilibria and the Aggregate Credit Supply Externality

A crucial mechanism behind most of normative results is the aggregate credit supply externality which arises due to imperfect competition in the banking sector. To illustrate this channel, we analyze a symmetric equilibrium with no aggregate uncertainty. Recall that the bank-level price rule is the following condition:

$$\frac{\mathrm{p(j)}}{\mathrm{P}} = \left[\underbrace{\frac{\theta}{\theta-1}}_{\text{Markup}} \underbrace{\frac{1}{\Big(1-\nu(\mathrm{j})\Big)\mathrm{R^T(j)} - \bar{\mathrm{R}}(\mathrm{j})} \frac{1}{\mathrm{P}} \mathrm{K}^{\zeta_2 - 1}}_{\text{Marginal Cost}} \right]^{\frac{1}{1+\theta(\zeta_2-1)}}$$

Now, the *aggregate* price rule can be determined if we solve for a symmetric, non-stochastic equilibrium. We require several additional simplifying assumptions. First, set $\xi(\mathrm{j})$ to the ergodic mean for all local markets (j). With no ex-post returns heterogeneity, the (representative) intermediary faces a homogenous probability of default and interest rate on deposits. That is, we work with the average bank from the distribution. The rate-setting rule for the aggregate economy is:

$$\bar{\mathrm{R}} = \Big(1-\nu\Big)\mathrm{R^T} - \frac{\theta}{\theta-1}\frac{1}{\mathrm{P}}\mathrm{K}^{\zeta_2-1} \tag{21}$$

The equation above is a downward-sloped demand curve for bank financing. The slope of the line is independent of the elasticity of substitution, which acts as a horizontal curve shifter. A symmetric equilibrium is achieved together with an the upward-sloped aggregate cost curve, which is the funding cost rule A.2 from the household's problem.:

$$\bar{\mathrm{R}} = \frac{1-\Lambda\upsilon \mathrm{x} \mathrm{R^T}}{\Big(1-\upsilon\Big)\Lambda} \tag{22}$$

It is straightforward to see that the rate is increasing in K because (a) $\bar{\mathrm{R}}$ is decreasing in the deposit recovery rate x and (b) x is decreasing in the leverage ratio ϕ. Everything else equal, as debt-financed capital grows, the recovery rate falls and the deposit rate goes up. In the symmetric equilibrium, the average rate $\bar{\mathrm{R}}$ is increasing in the supply of capital K. Jointly, equations 21 and 22 determine the credit market outcome. Equilibrium exists for a given $\zeta_2 > 1$. Decrease in θ shifts the demand schedule leftwards.

The proposition below summarizes these results and establishes the credit supply externality:

Proposition 2 (Aggregate Credit Supply Externality). *For any finite $\theta > 1$, lower θ yields under-utilization of aggregate capital in the symmetric equilibrium.*

Proof: Appendix A

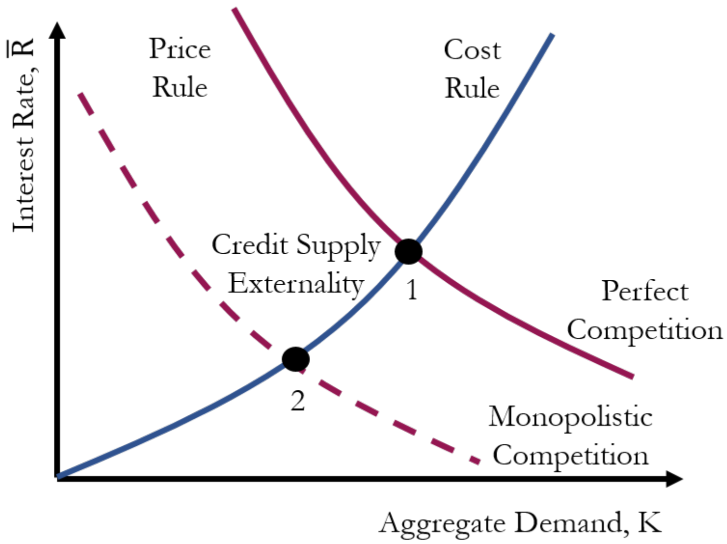
To start, consider the case of perfect competition in the credit market and set $\theta \to \infty$. Now, as a finite but large θ falls, the wedge between the perfect competition benchmark and the market outcome widens. Figure 3 visualizes this point for a given $\zeta_2 > 1$. It plots the credit demand schedule under perfect (PC) and monopolistic (MC) competition assumptions. On the horizontal axis we have aggregate demand K. Prices (the equilibrium deposit rate) are on the vertical axis.

The transition from PC to MC, driven by a decline in θ, is associated with the widening of the credit markup wedge and an increase in the social deadweight loss from monopolistic competition in lending. Credit market power leads to an *aggregate credit supply externality*. The MC equilibrium is associated with under-lending and under-utilization of risky capital relative to the PC benchmark. Banks do not internalize the effect of the truple $\{\mathrm{k(j)}, \mathrm{d(j)}, \mathrm{p(j)}\}$ on the aggregate price of capital. Higher private margins cause the average price to rise, which in turn pushes down the equilibrium return to investment for all banks. This result mimics the canonical aggregate demand externality in Blanchard and Kiyotaki (1987) and Farhi and Werning (2016). Our approach is different as our externality arises in the credit market on the asset side of banks balance sheets, and we do not rely on nominal rigitidies. This is a separate channel of transmission from the financial sector to the real economy.

2.12 Additional Results

In the Appendix, I explore several additional results and channels. First, in Section A.3, we discuss asset pricing implications in the model. Second, in Section A.4 we prove complementarity of book leverage and bank net worth, a result that is consistent with the empirical evidence in Coimbra and Rey (2019) and Gopinath et al. (2017). Finally, in Section A.5 we show that the model can generate size-dependent sensitivity to cost of funds shocks, in line with the evidence in the seminal contribution of Kashyap and Stein (2000).

Figure 3: **Monopolistic Credit Market Equilibrium**



Notes: Visualization of the static, symmetric equilibrium with monopolistic and perfect competition in bank lending.

3 Taking the Model to the Data

In this section I discuss how I quantify the model and map it to the data. I begin by discussing the calibration approach. Next, I show predictions of the model on cross-sectional bank behavior and define the Marginal Propensity to Lend.

3.1 Calibration

Table 1 lists all the chosen parameters. The model period is one quarter. We begin by describing standard macro parameters. We set the share of aggregate capital in production $\alpha = 0.36$. The discount factor β =0.996 is chosen to target a steady-state annual risk-free rate of 2.60%. We assume log-utility in consumption ($\sigma_{\text{h}} = 1$).

For parameters in the banking block, we set the dividend payout ratio $\sigma = 0.97$ which is broadly consistent with the exit rate of banks in the U.S. According to the FDIC, there were roughly 11000 commercial banks in the U.S. in 1980. This number has dropped to 5000 by 2019. This implies an average annual exit rate of 3% and a life expectancy of a banker of about 8.25 years. The fraction of divertible assets $\lambda = 0.12$ targets a steady state bank leverage ratio of roughly 10. Endowment of new entrants and the fixed cost of entry are calibrated in order to keep the entry rate at around 5%. Parameters that govern non-interest variable and fixed expenses (ζ_1, ζ_2, FC) are chosen to be consistent with empirical evidence on increasing returns in banking (Wheelock, 2011).

We calibrate θ with the reduced-form estimate of the credit demand elasticity from Jamilov

Table 1: **Model Parameters**

Parameter	Description	Value
	Macro	
α	Share of capital in production	0.36
β_{h}	Impatience	0.996
σ_{h}	Household risk aversion	1
	Banking	
σ	Dividend payout ratio	0.97
λ	Share of divertible assets	0.12
n_0	Entry starting endowment	30% of N
e	Entry fixed cost	0.11
$\frac{1}{\zeta_1}$	Monitoring cost linear	0.01
ζ_2	Monitoring cost quadratic	1.18
	Monopolistic Banking Block (Jamilov '20)	
θ	Credit Demand Elasticity	1.2
	Idiosyncratic Risk (Galaasen et al. '20)	
κ	Fraction of wealth exposed to idiosyncratic risk	0.5
ρ_ξ	Serial correlation of idiosyncratic risk	0.529
σ_ξ	SD of idiosyncratic risk	0.074

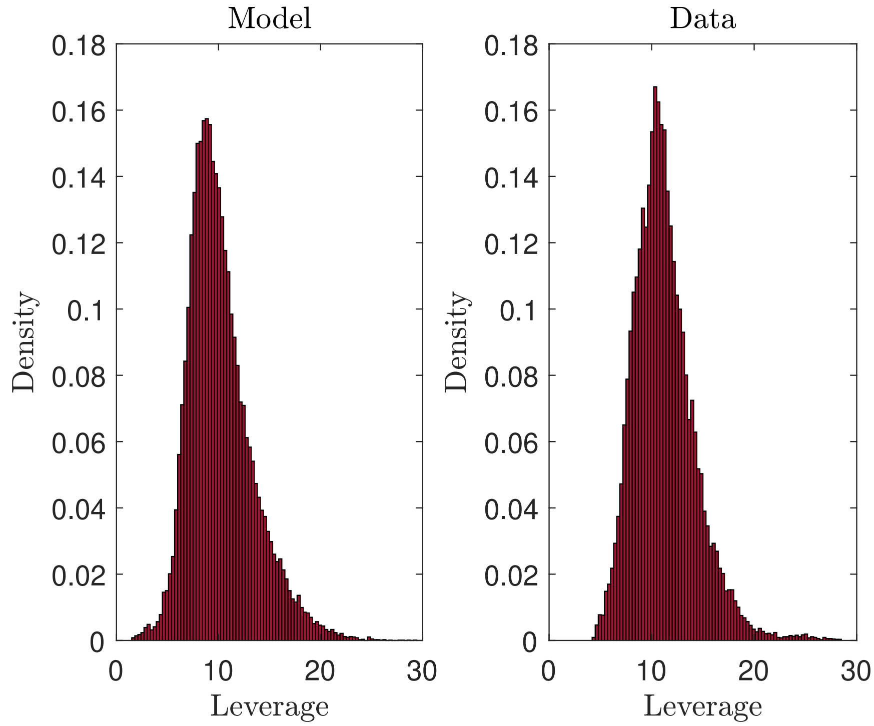
Notes: This table summarizes chosen model parameters.

(2020b) and set it to 1.2. This implies an average, constant credit markup of 6 over the marginal cost, which is broadly in line with the Corbae and D'Erasimo (2019) evidence on markups and margins in the financial industry, particularly of large commercial banks.

Calibration of the idiosyncratic return process follows closely Galaasen et al. (2020). Galaasen et al. (2020) estimate the pass-through of granular firm shocks on bank-level returns using matched bank-firm data from Norway. They fit a linear fixed effects model with AR(1) error disturbances into the aggregated granular credit risk process and estimate annual persistence and standard deviation parameters which correspond closely to our chosen values for quarterly periodicity. The fraction of financial wealth that is exposed to idiosyncratic risk $\kappa = 0.5$ is consistent with the pre-financial crisis share of the shadow banking business in overall banking in the U.S. (Gorton and Metrick, 2010). The idiosyncratic risk process is discretized with the Tauchen (1986) method.

In Appendix D we discuss in detail the numerical algorithm used to solve the model.

Figure 4: **Distribution of Bank Leverage - Model Meets Data**



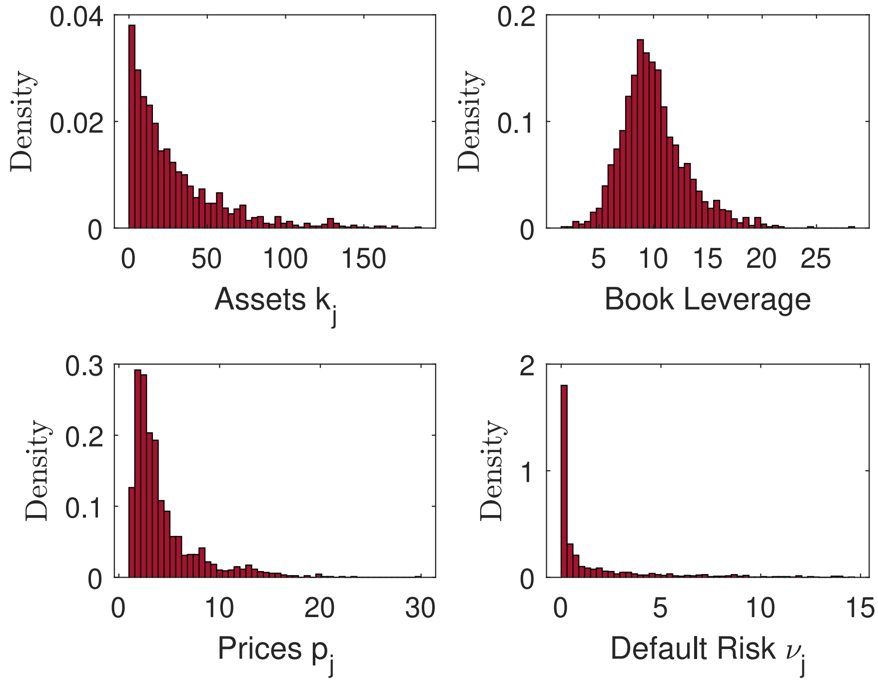
Notes: Distribution of book bank leverage in the model and in the data. Data source: Compustat Banks. Sample is restricted to U.S. commercial banks. Bank-level leverage ratios are pooled over all banks and quarters over the 1980:q1-2020:q1 period.

3.2 Matching the Distribution of Bank Leverage

Our calibration target is the cross-sectional distribution of bank (book) leverage. I collect bank-level total assets and total equity information from Compustat. The sample is restricted to consolidated U.S. commercial banks. Book leverage is defined as total assets over total equity. The distribution of book leverage is pooled over all banks and years over the 1980:q1-2020q1 period and truncated at the 1% and 99% levels. The first three moments of the resulting distribution are 11.22, 3.01, and 0.98, respectively. That is, the average leverage ratio is roughly 11, and the distribution is disperse and right-skewed.

The calibration approach entails searching for the set of parameters $\{\rho_\xi, \sigma_\xi\}$, within the ballpark of reduced-form estimates in Galaasen et al. (2020) and Jamilov (2020b), such that the model-generated stationary distribution of book leverage matches the data exactly. Figure 4 plots the result of the calibration. The model-generated distribution moments are 11.11, 3.40, and 1.01, respectively. The level of leverage responds best to σ_ξ, volatility of the idiosyncratic process, as well as to the fraction of divertible assets λ. Skewness of the distribution can be tweaked with persistence of idiosyncratic shocks ρ_ξ.

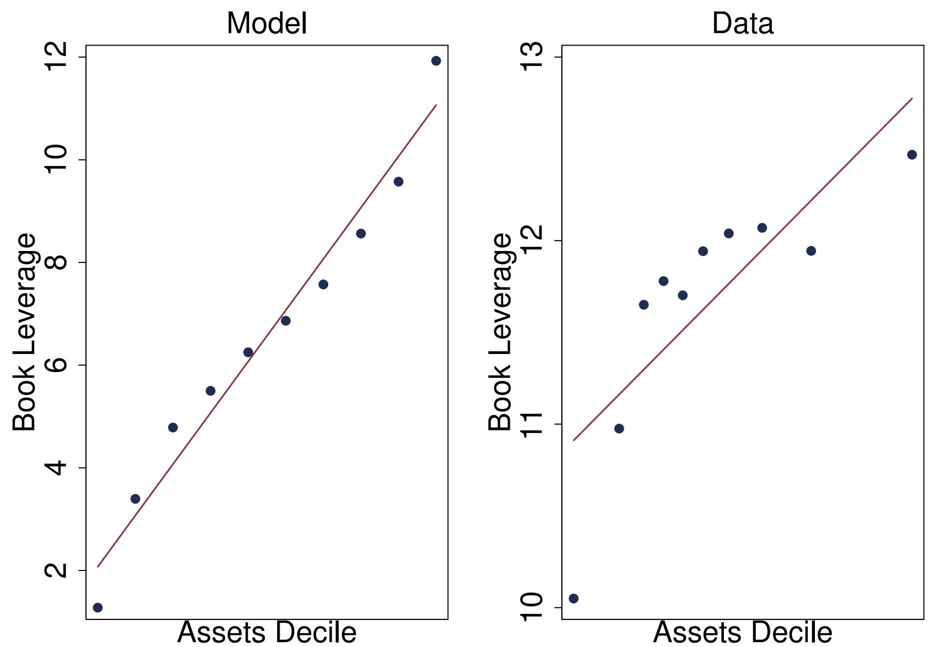
Figure 5: **Stationary Bank Distributions in the Model**



Notes: Model-generated stationary distributions of assets, book leverage, relative prices, and default risk probability.

The next step is to analyze the behavior of untargeted cross-sectional distributions. Figure 5 plots the histograms of stationary distributions of bank assets k(j), book leverage, relative prices p(j) and default risk ν(j). The model generates a right-skewed distribution of bank size, as proxied by total assets. There is a small fraction of very large intermediaries which, as we will see below, also have high book leverage ratios. Distributions of relative prices and default risk are also very right-skewed: the vast majority of banks in the distribution are insulated from default risk considerations. However, a small fraction of low-net worth institutions are at a high risk of default. For those same institutions, high risk feeds into higher marginal costs and relative prices. The right tails of the distributions of p(j) and ν(j) are driven by the same mass of low-n(j), low-R^T(j) intermediaries.

Figure 6: **Bank Leverage and Total Assets**



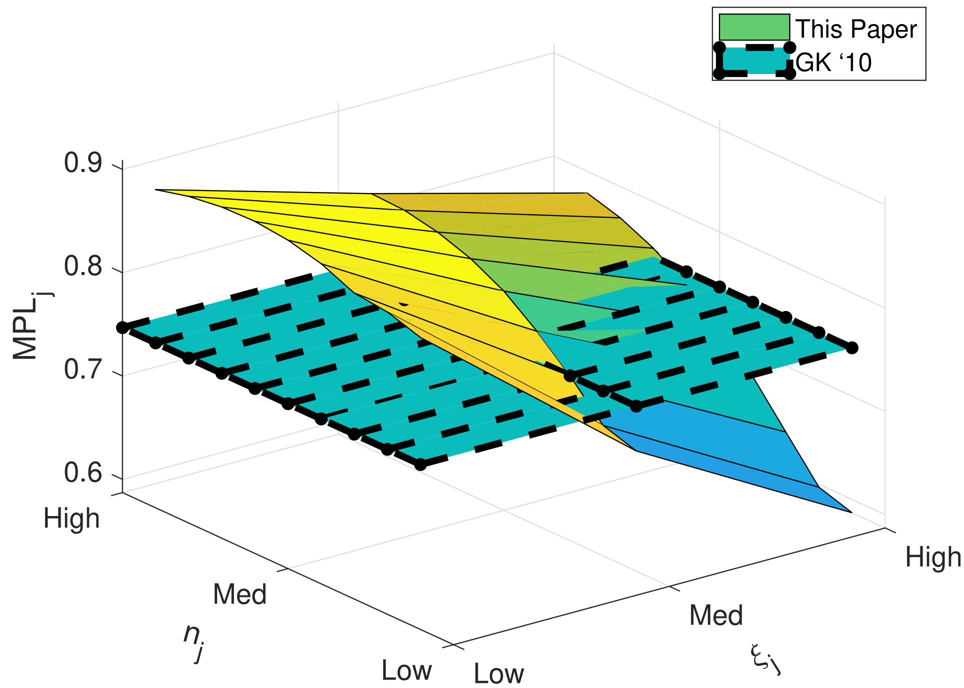
Notes: Correlation of bank total assets and book leverage, in the model and data. Data source: Compustat Banks. Sample is restricted to U.S. commercial banks. Data is pooled over all banks and quarters over the 1980:q1-2020:q1 period and binned into 10 deciles of the total assets distribution. For each bin, book leverage is computed as the ratio of bin-specific total assets over bin-specific total equity. The same binning procedure is applied to the model.

We conclude this section by showing how, in line with the data, the model generates a positive correlation between bank book leverage and size in the cross section. We proxy bank size with total assets, similarly to Adrian and Shin (2010) and Coimbra and Rey (2019), and continue with the same sample of banks from Compustat as before. In order to minimize the influence of outliers and noise, we plot binned scatter plots. Specifically, we construct ten deciles (bins) of the pooled distribution of total assets. For each bin, we compute the ratio of bin-specific total assets over bin-specific total equity. Both in book values, as before. The same procedure is performed on model-generated data. Figure 6 plots the outcome. The positive relationship arises in the model because marginal costs decline with bank size and larger intermediaries can extend more loans at a lower cost of external funds.

3.3 Marginal Propensity to Lend Heterogeneity

An essential endogenous object that the model generates is the distribution of the Marginal Propensity to Lend (MPL). At the level of a bank, MPL(j) is defined as the elasticity of assets k(j) to changes in net worth n(j). Banks of different levels of n(j) have ex-post heterogeneous marginal

Figure 7: **Marginal Propensity to Lend**



Notes: MPL heterogeneity in the baseline economy and in Gertler and Kiyotaki (2010).

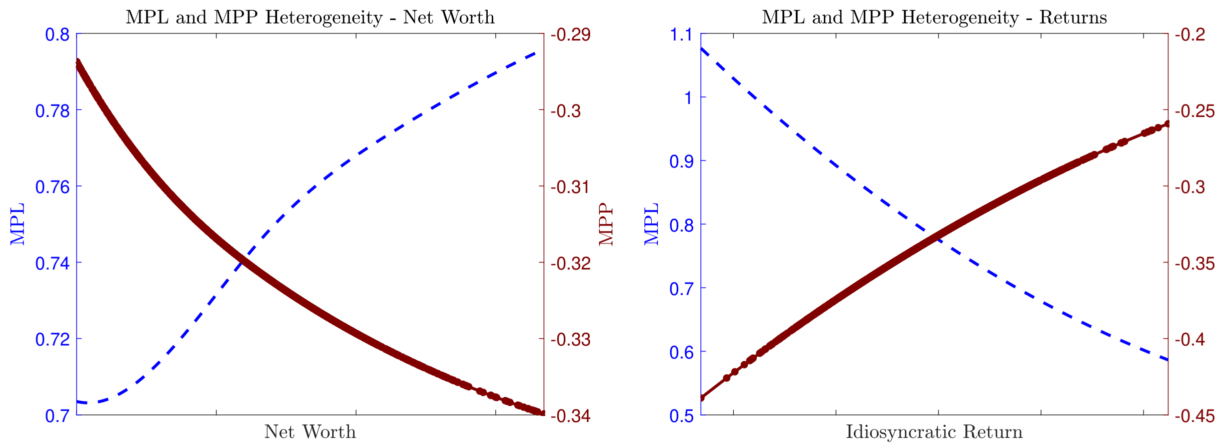
costs due to variability in default risk, deposit rates, and idiosyncratic portfolio returns. Different banks, therefore, have heterogeneous elasticities with respect to the same exogenous change in net worth. Integrating over the joint distribution of $\{\text{n(j)}, \xi\text{(j)}\}$ grants the economywide aggregate MPL. With persistent heterogeneity, the aggregate MPL does not correspond to the MPL(j) of the average, "representative" intermediary.

Our ultimate object of interest is the *macro elasticity* of aggregate output/demand with respect to shocks to the financial sector. Consider an aggregate shock or policy intervention such that every bank in the distribution receives an additional unit of net worth. Formally, the elasticity with respect to systematic net worth fluctuations is:

$$\frac{\partial \mathrm{Y}}{\partial \mathrm{N}} = \underbrace{\frac{\partial \mathrm{Y}}{\partial \mathrm{K}}}_{\text{MPK}} \times \underbrace{\int_{\mathbf{B}} \frac{\partial \mathrm{k(j)}}{\partial \mathrm{n(j)}} \mu(\mathrm{dn}, \mathrm{d}\xi)}_{\text{MPL(j)}} \tag{23}$$

Where the first component is the model-specific marginal product of capital and the second component is the distribution of MPL(j) integrated over all banks in the economy. Generally, both the macro elasticity (left-hand side of the equation) and each of the two right-hand side components are endogenous and depend on calibration. We will explore how different assumptions in the model

Figure 8: **MPL and MPP Heterogeneity**



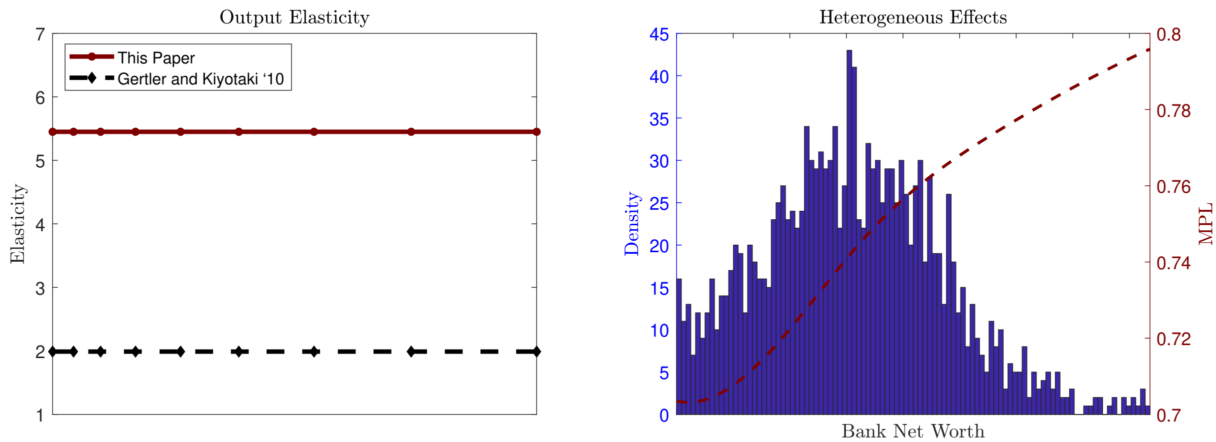
Notes: MPL and MPP heterogeneity in the baseline economy.

affect MPL heterogeneity and the macro elasticities in the next section.

Given the baseline calibration from Section 3.1, Figure 7 plots MPL(j) heterogeneity from the stationary industry equilibrium. The figure contrasts MPL heterogeneity in this paper with the Gertler and Kiyotaki (2010) economy without imperfect competition and idiosyncratic risk. There are three core takeaways from the figure. First, MPL(j) is increasing in bank net worth. This is a general result that, as we will see in the next section, is robust to most modelling assumptions and calibration scenarios. Second, MPL(j) decreases with idiosyncratic returns. This result typically depends on calibration of ζ_2 and, more broadly, on the degree of economies of scale in banking. We will return to this point in the next section. Finally, in the GK counterfactual, the MPL distribution is flat and essentially corresponds to the MPL of the representative intermediary. Depending on the extensive margin and the relative shares of very large and very small banks in the distribution, heterogeneity could be very important and the quantitative departure from GK very large.

MPL heterogeneity is intricately related to the *Marginal Propensity to Price* (MPP), defined analogously to the MPL as the elasticity of bank-level relative prices p(j) with respect to shocks to bank net worth n(j). We illustrate MPP heterogeneity in Figure 8. The left and right panels plot the dependencies of MPL(j) and MPP(j) on net worth n(j) and idiosyncratic risk ξ(j), respectively. We see that MPP(j) and MPL(j) are inversely related, which is due to the downward sloping demand function. Both objects are linked and reinforced through heterogeneous marginal costs, which are decreasing in bank net worth. MPP heterogeneity will be essential to understand the workings of direct lending facilities in the next section.

Figure 9: **Macroeconomic Effects of Systematic Equity Injections**



Notes: Response of aggregate output to a systematic, economywide bank equity injection.

4 Credit Policy

In this section we study the main quantitative experiment in the model - systematic and targeted credit policies. We consider four distinct policies: equity injections, direct lending facilities, liquidity facilities, and deposit guarantee schemes. We contrast sector-wide, systematic credit policies with a scenario where the government can potentially manage *any* bank in the distribution, an operation we label as *targeted credit policy*. We abstract from any operational costs of these credit policies and assume that they are funded by non-distortionary lump-sum taxes.

4.1 Systematic Equity Injections

We start with systematic equity injections where each bank in the distribution receives one additional unit of net worth at no marginal cost to society. We are interested in quantifying the impact of this policy on aggregate output. Measurement proceeds in two steps. First, we compute the distribution of MPL, again for our baseline economy and for the GK benchmark. Second, we compute the macro elasticity separately for each economy as per equation 4.1.

$$\frac{\partial \mathrm{Y}}{\partial \mathrm{N}} = \underbrace{\frac{\partial \mathrm{Y}}{\partial \mathrm{K}}}_{\text{MPK}=0.0734} \times \underbrace{\int_{\mathbf{B}} \frac{\partial \mathrm{k(j)}}{\partial \mathrm{n(j)}} \mu(\mathrm{dn}, \mathrm{d}\xi)}_{\text{MPL}=0.7428}$$

The result is plotted on Figure 9. On the right panel, we plot the ergodic density of bank net worth overlaid with the MPL schedule. On the left panel, we plot the elasticity of aggregate output with respect to systematic equity injections. In the baseline economy with heterogeneity, the macro elasticity is 5.45% and the aggregate MPL is 0.74.[17] That is, recapitalization of the whole banking sector by 10% raises aggregate output by 54.5 basis points. The elasticity is twice as large as in the GK benchmark. This is driven by two factors. First, banks are heterogeneous across the intensive margin in terms of the MPL(j). Second, the stationary distribution of bank net worth is not clustered around the "average" bank. In other words, there is significant dispersion in ex-post measures of size, returns, and risk. We have thus shown that bank heterogeneity matters explicitly for macroeconomic aggregates and for the design of even systematic, economywide policies.

4.2 Targeted Equity Injections

In this section we move beyond systematic credit policy analysis of Gertler and Kiyotaki (2010) and estimate conditional macro elasticities when equity injections are allowed on any *individual* bank in the distribution. We proceed in three steps. First, we break the distribution of bank net worth into ten bins (deciles). For each decile $\iota = 1 \dots 10$, we assume that the government injects one unit of equity to every bank in ι but not to anywhere else in the distribution. Second, we re-evaluate the cross-sectional distribution after the policy shock. Third, we compute the macro elasticity with respect to targeted policies using the same, equilibrium MPL distribution but integrating it over different ex-post distributions of bank net worth after the equity injections took place. We thus run ten separate experiments, one per each decile of the size distribution, and compute the conditional impact on aggregate output ten times.

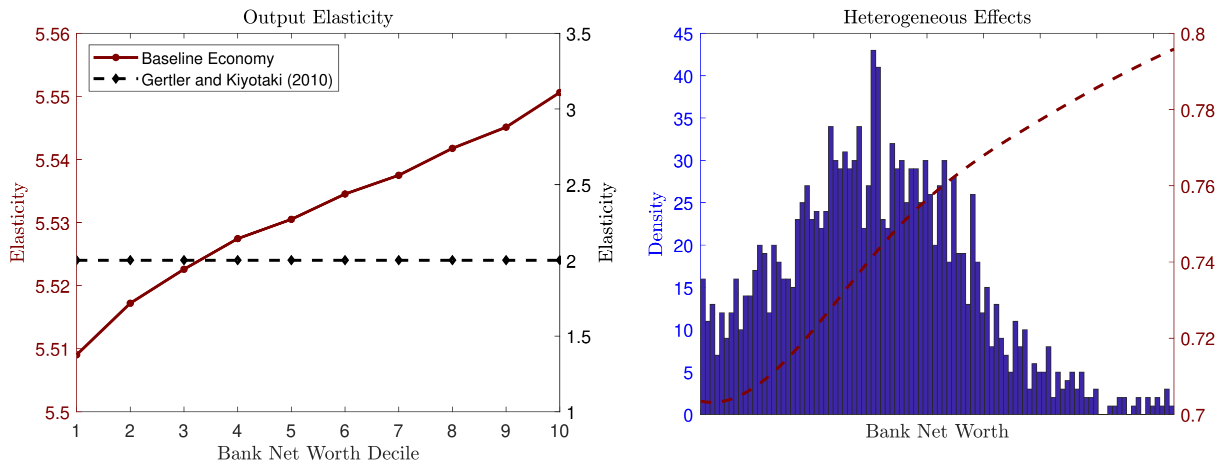
Figure 10 plots the result. We observe that there are efficiency gains from injecting equity into large intermediaries. The elasticity of aggregate output with respect to decile-specific credit policies is an upward-sloping line. This result is driven by the shape of the MPL distribution - larger banks have a greater equilibrium MPL, which is in turn due to big banks having lower marginal costs and relative prices. Abstracting from any normative implications or second-level effects on financial stability or systemic risk, if the objective of the government is purely to stimulate aggregate lending and demand, then "bailing out" big banks yields a bigger bang for the buck.[18] Note that in contrast to the baseline economy with heterogeneity, the macro elasticity in the GK economy is flat at roughly 2% and invariant to bank size because of the absence of ex-post returns heterogeneity and scale dependency.

In the rest of this section, we explore how different mechanisms in the model affect MPL heterogeneity and the macro elasticity. We focus on the role of decreasing vs increasing returns to scale in banking, imperfect competition, and idiosyncratic risk.

[17] How sensible are our estimates of the MPL? A comparison could be made with the empirical banking studies that quantify bank loan supply elasticities. For example, the seminal contribution of Khwaja and Mian (2008) finds that the elasticity of bank lending to unexpected liquidity shocks is 0.6, which is in the ballpark with our estimate of 0.74.

[18] For simplicity, we assume that targeted credit policies or bailouts are unexpected and do not generate additional moral hazard frictions ex-ante.

Figure 10: **Macroeconomic Effects of Targeted Equity Injections**



Notes: Responses of aggregate output to targeted, decile-specific bank equity injections.

The Role of Increasing vs Decreasing Returns to Scale In the baseline calibration, banks face non-interest expenses that are convex in assets under management.[19] We now explore the opposite case, that is when $\zeta_2 < 1$ and non-interest costs are concave in bank assets. Figure 11 reports the result in the usual format. The right panel plots the stationary distribution of net worth and the MPL schedule. Notice how the distribution is clustered around the right tail as a larger fraction of intermediaries now manages to accumulate a lot of net worth thanks to strong increasing returns to scale. Note how the level of the MPL distribution is significantly higher than in the baseline economy.

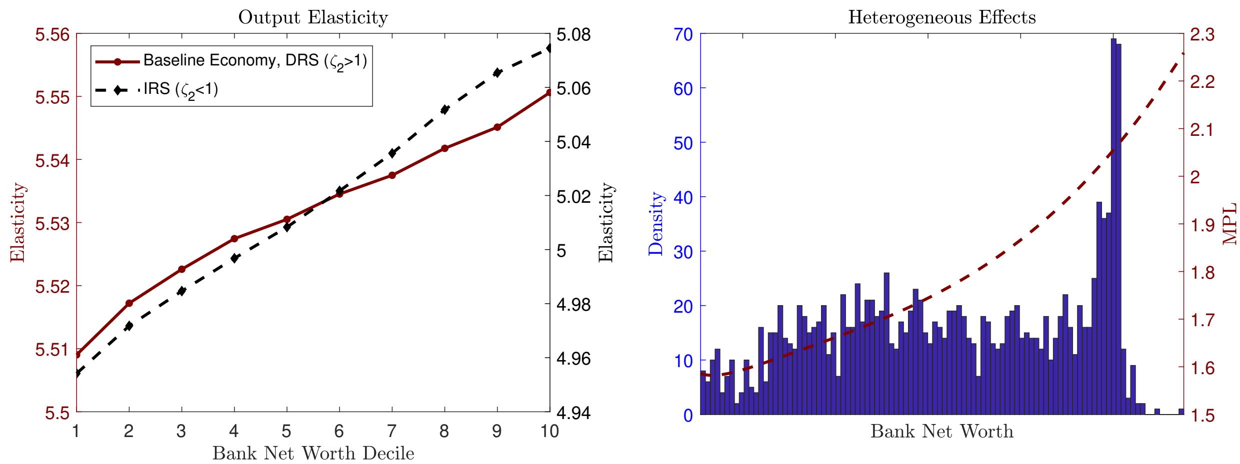
The left panel of the figure plots macro elasticities for the economies with increasing and decreasing (baseline) returns. The slope of the output elasticity schedule is now more steep - relative efficiency gains from injecting equity into large vs small banks are more pronounced. This result is intuitive, since with increasing returns the slope of the total cost function is now steeper (more downward-sloped). Quantitatively, aggregate output increases by more than 12 basis points if policy targets the top vs the bottom size decile, or roughly \$245 billion in 2020 U.S. real GDP equivalents.[20]

The Role of Idiosyncratic Risk We now comment on how idiosyncratic risk contributes to the equity injections experiment. We set $\kappa = 0, \sigma_\xi = 0$, which shuts down ex-post heterogeneity in

[19]Despite this channel, *total* cost of operations (interest plus non-interest costs) still declines in bank size and the financial sector exhibits increasing returns to scale, in line with the empirical evidence Wheelock (2011). Adding fixed operational costs amplifies increasing returns but is not otherwise necessary.

[20]Increasing returns to scale appears to be the empirically more realistic assumption. However, we required $\zeta_2 > 1$ for some of our prior analytical results, particularly existence of a *symmetric* equilibrium. For this reason of internal completeness, we use decreasing returns as our baseline case.

Figure 11: **Targeted Equity Injections and Economies of Scale**



Notes: Responses of aggregate output to targeted, decile-specific bank equity injections for different parametrizations of scale economies.

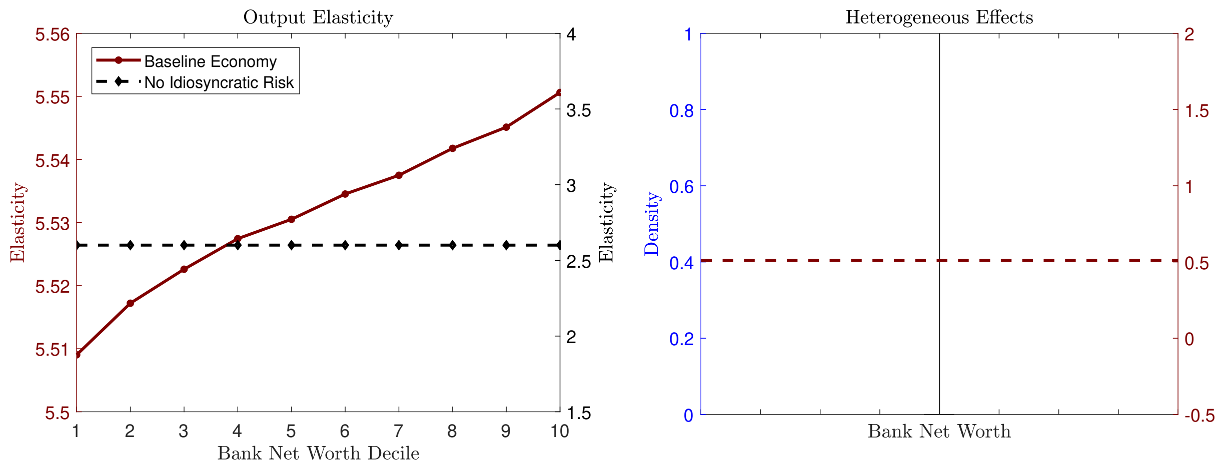
returns. This, in turn, leads to no ex-post heterogeneity in any quantity or price variables in the cross-section. The banking sector reduces to the case of a representative intermediary. Recall that we still retain the monopolistic competition assumption so that the bank charges a markup over the homogenous marginal cost.

Figure 12 plots the results. We see that without idiosyncratic risk, the distribution of bank net worth has no dispersion as all banks in the distribution are the same ex-post. Similarly, the MPL(j) schedule is flat and corresponds to the MPL of the representative bank. On the left panel, we see that the macro elasticity is the same regardless of which decile of the bank net worth distribution gets targeted. The level of the macro elasticity is greater than in the Gertler and Kiyotaki (2010) economy from Figure 9. This is because the economy still features monopolistic competition, which yields under-lending in equilibrium and a higher marginal return on aggregate capital. All in all, we conclude that uninsurable idiosyncratic return risk is crucial for the analysis of targeted policies. This is the only source of ex-post heterogeneity in the model, and it's necessary that idiosyncratic shocks ξ(j) are at least partially uninsured.

The Role of Monopolistic Competition We conclude this section by estimating the impact of monopolistic competition in bank lending on the macro elasticity of direct equity injections. Recall from Equation 2.5 that as $\theta \longrightarrow \infty$, the distribution of relative prices collapses to unity. However, because of non-linear non-interest expenses and presence of idiosyncratic return shocks, there is still ex-post heterogeneity in bank assets, net worth, leverage, etc. The only difference is that now the values of market and book leverage equalize in the cross section.

Figure 13 plots the result from the equity injections experiment for the case of perfect compe-

Figure 12: **Targeted Equity Injections with No Idiosyncratic Risk**



Notes: Responses of aggregate output to targeted, decile-specific bank equity injections for the baseline economy with and without idiosyncratic risk and incomplete markets

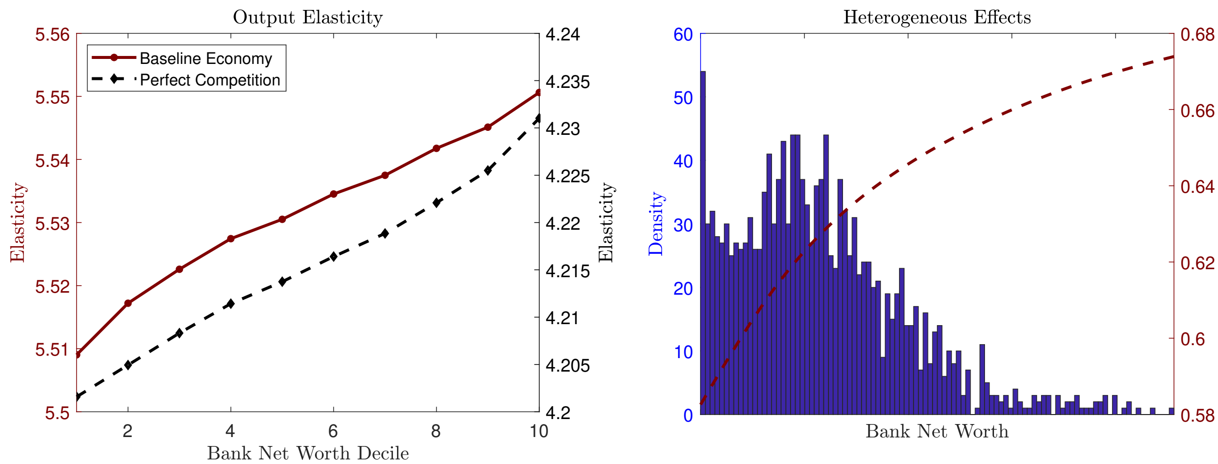
tition. On the right panel, we see that the MPL schedule is increasing in bank size the same way as before. On the left panel, we compare conditional macro elasticities in the perfect- and imperfect-competition economies. The two curves are both upward-sloped, with the average elasticity in the perfect-competition case being one percentage point below than in the baseline. This occurs because imperfect competition induces equilibrium under-lending and a higher average systematic return on aggregate capital. Overall, we reach the same conclusion that it's efficient to inject equity into large intermediaries. Even when banks have no market power.

4.3 Targeted Direct Lending Facility

In this section we study an alternative form of targeted credit policies - a direct lending facility. Broadly speaking, we are referring to a scenario where the monetary authority takes over market lending on behalf of the intermediaries. In the model, this corresponds to the market for differentiated capital goods. Crucially, the policy alters the distribution of marginal costs in the banking sector. We assume that the cost of funds of the central bank is lower than of any bank in the ergodic distribution. Effectively, the central bank is offering funds at a subsidized price. Firms that borrow from banks with ex-ante high marginal cost are the biggest beneficiaries of this policy. The central bank is not balance sheet constrained and doesn't face any moral hazard frictions. We also assume that there is no crowding out of existing lending by market participants.

We consider *targeted direct lending*, and run decile-specific policy counterfactuals, similarly to the way we approached direct equity injections. We assume that the central bank takes over market lending of individual banks of the distribution. Our objective, as before, is to compute conditional

Figure 13: **Targeted Equity Injections with Perfect Competition**



Notes: Responses of aggregate output to targeted, decile-specific bank equity injections for the baseline economy with and without monopolistic competition

elasticities of aggregate output with respect to targeted policy interventions. Formally, we compute the following object in the model:

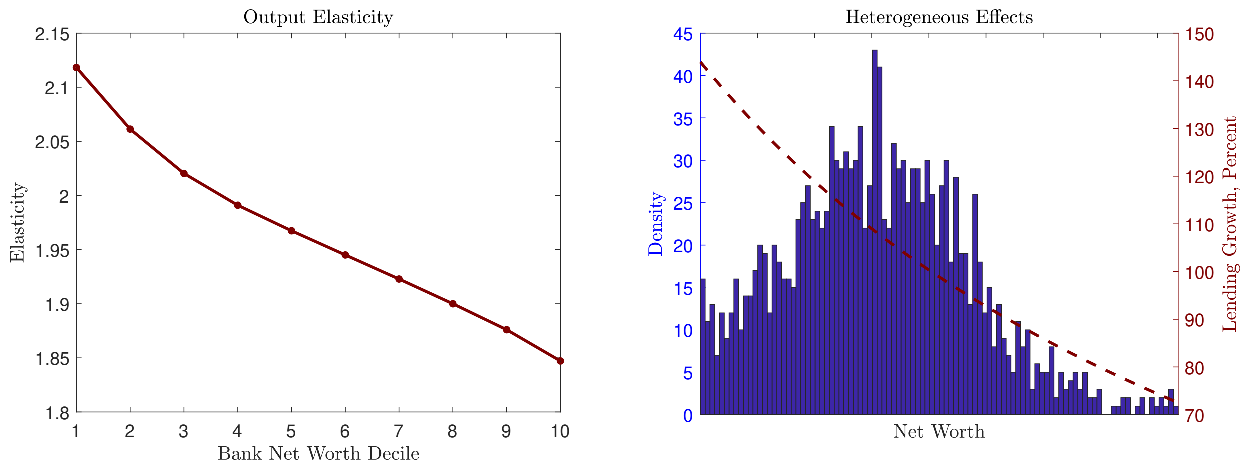
$$\frac{\partial Y}{\partial \hat{P}} = \frac{\partial Y}{\partial \hat{K}} \times \int_{\mathbf{B}} \frac{\partial k(j)}{\partial \hat{p}(j)} \mu(dn, d\xi) \tag{24}$$

where the left-hand side is the response of aggregate output to the direct lending facility introduction, captured by $\hat{P}$. The distribution of $\hat{p}(j)$ reflects new relative prices after the central bank's takeover of a particular institution (j). Specifically, for each decile of the distribution, we assume that the new relative price $\hat{p}(j)$ is consistent with lowest marginal cost available in the state space.[21] On the right-hand side we show the decomposition into the marginal product of capital (which is evaluated at the new, post-policy level of capital) and the distribution of bank-level changes in lending with respect to the new, central bank-generated relative prices. The central bank is thus essentially adjusting the market's credit supply elasticity.

Results of this policy experiment are plotted on Figure 14. On the right panel we see the stationary distribution of net worth from the baseline economy overlaid with the bank-level growth in lending in response to the policy intervention. Notice how the schedule is *decreasing* in bank net worth, because the marginal cost is lower for larger banks. The direct lending policy scheme generates a relative price advantage only for the low-net worth institutions whose cost of funds is high ex-ante. This is in direct contrast to the effects of direct equity injections from the previous

[21] Technically, the state space is larger than the ergodic distribution. For this reason, even the largest banks in the distribution increase lending in response to this policy. Alternatively, we can normalize the $\hat{p}(j)$ to be consistent with the lowest marginal cost in the ergodic distribution. Results and conclusions would not change.

Figure 14: **Macroeconomic Effects of Targeted Direct Lending Facilities**



Notes: Responses of aggregate output to targeted, decile-specific direct lending policy.

section. On the left panel of the Figure, we compute the elasticity of aggregate output with respect to decile-specific direct lending intervention, which is intutitively downward-sloping. Direct lending interventions targeted at smaller institutions are more effective at stimulating lending and demand growth.

4.4 Targeted Liquidity Facilities

Financial crises are typically associated with tightening of liquidity constraints. As opposed to the lack of credit worthiness of borrowers, it is the lack of liquidity on the credit supply side that contributes to rising excess returns. In our model, banks face a liquidity constraint in the form of the moral hazard-induced cap on leverage-taking. The fraction of divertible assets - λ - controls the degree of constraint tightness and is generally part of the exogenous environment. We now suppose that the government can step in and augment λ on behalf of private lenders. In particular, we allow λ to be relaxed on *any* bank in the distribution.

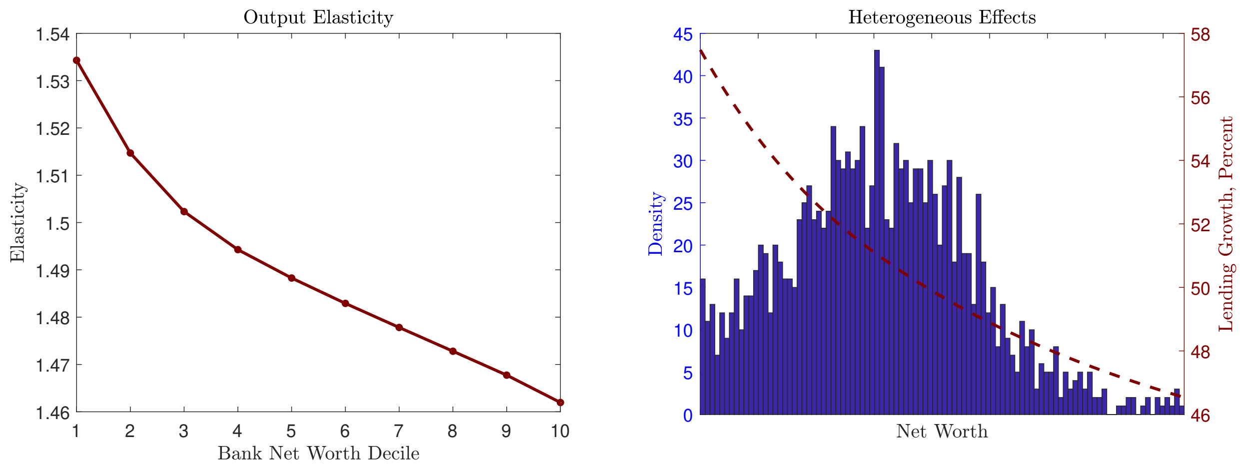
In order to facilitate the cleanest possible analysis, we assume that the leverage constraint binds on all banks in the distribution.[22] With the binding leverage constraint, it is straightforward to solve for the bank-specific leverage ratio:

$$\phi(\mathrm{j}) = \frac{\nu_{\mathrm{a}}(\mathrm{j})}{\lambda - \mu_{\mathrm{a}}(\mathrm{j})} \tag{25}$$

where, as before, $\phi(\mathrm{j})$ is market leverage, $\nu_{\mathrm{a}}(\mathrm{j})$ is the discounted cost of bank liabilities, $\mu_{\mathrm{a}}(\mathrm{j})$ are excess returns on the risky asset. Notice how according to this formula, relaxation of liquidity

[22]This is a realistic assumption given that these types of policies are usually only implemented in crisis episodes, precisely when liquidity and leverage constraints of market lenders tighten.

Figure 15: **Macroeconomic Effects of Targeted Liquidity Facilities**



Notes: Responses of aggregate output to targeted, decile-specific liquidity facilities.

conditions (as proxied by a reduction in λ) increases banks appetite for leverage. Everything else equal, this raises credit supply in the market.

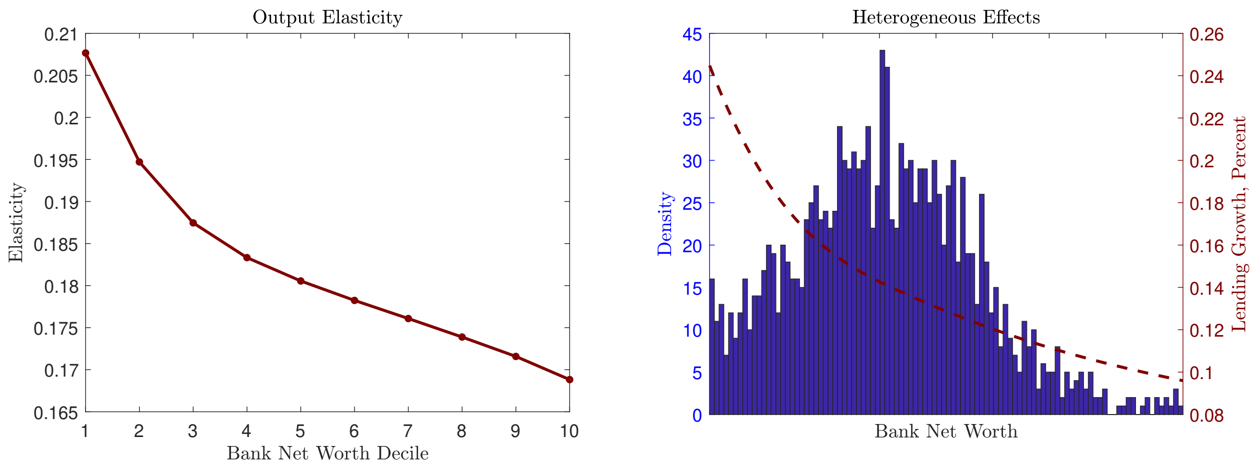
We proceed by assuming that the government intervenes by lowering λ_ι on decile $\iota = 1 \ldots 10$ of the banks net worth distribution by 25% relative to the baseline value of 0.12. The exogenous shock is thus invariant to the region of the distribution which is targeted. The only variant in this policy intervention is the decile of the bank net worth distribution. For each of the ten policy counterfactuals, we compute the MPL(j) distribution and the output elasticity the same way as before.

Figure 15 presents the result. We see that the differential impact of this policy is concentrated in the left tail of the distribution - smaller banks increase their credit by more. On the left panel we see how this translates into a downward-sloped output elasticity curve. This result arises because the marginal effect of λ_ι on ϕ_ι is negative and declining with bank size due to diminishing marginal costs of funds. Because we assumed that the constraint is always binding, we have therefore isolated the intensive margin of the total effect. Suppose now that the constraint can bind occasionally. The Lagrange multiplier on the constraint declines in bank size. In other words, the constraint is generally slack for big banks and binding for small banks. Relaxation of the moral hazard friction is therefore much more likely to differentially benefit small banks even if we allow the constraint to bind occasionally. Results of Figure 15 would therefore not change.

4.5 Bank-Level Debt Guarantees

Our baseline economy has no deposit insurance or any form of explicit/implicit debt guarantee schemes. The distribution of bank-level default risk ν(j) is therefore priced competitively into the

Figure 16: **Macroeconomic Effects of Deposit Guarantees**



Notes: Responses of aggregate output to targeted, decile-specific bank equity injections for the baseline economy with and without deposit insurance.

distribution of bank deposit rates $\bar{R}(j)$. In practice, however, deposit insurance is a perennial feature of traditional banking (Farhi and Tirole, 2017). We now introduce deposit insurance transfers that compensate the household for whatever losses any given bank in the distribution could deliver ex-post. In other words, if n(j) turns negative and the realized deposit recovery rate is less than unity, the government steps in on behalf of the bank and repays the promised volume of deposits in full. We assume that there are no coverage limits and that transfers are funded by lump-sum taxes on the household ex-post.

In the model, introduction of deposit guarantees shuts down two sources of marginal cost heterogeneity - deposit rates and default risk. That is, there is an explicit guarantee that no bank in the distribution can ever default. This immediately implies that there is no deposit rate heterogeneity as all banks can borrow at the risk-free rate - there is no deposit spread. Recall that as long as markets are incomplete, banks still face idiosyncratic returns $\xi(j)$, which feeds into balance sheet heterogeneity ex-post. Just as before, we segment the distribution of bank net worth into 10 deciles $\iota = 1 \ldots 10$. For each ι, we shut down the idiosyncratic default risk channel, which lowers the cost of funds of all banks in that decile to the risk-free rate. We then compute the MPL(j) and the macro elasticities subject to the new conditional distributions ten separate times, one per each affected decile.

Figure 16 presents the result. The right panel shows how banks of different levels of net worth respond to the introduction of the debt guarantee scheme. The left panel plots the macro elasticity schedule. We see from both panels that deposit guarantees increase bank risk-taking and credit supply, which leads to output growth. Quantitatively, the level effect is rather small, if compared to other three credit policy types, but generally depends on the calibration of σ_ξ.

In terms of heterogeneous effects, deposit guarantees increase lending of small banks by more. The distribution of bank default risk is concentrated in the left tail of the net worth density. Elimination of this channel of risk therefore mechanically affects only the smallest intermediaries, which explains the slope of the curve. Not only are the small banks responsible for fewer assets under management, but their MPL(j) is relatively low too, which contributes to the level effect being limited. Interestingly, deposit insurance does not materially increase overall risk-taking in general equilibrium: interest expenses are already close to the lower bound (the risk-free rate) for the majority of incumbent banks.

We summarize our results so far. Bank heterogeneity matters for the design of credit policies. Even if policies are systematic and bank size-invariant, the distribution has implications on the elasticity of aggregate demand with respect to the policy intervention. The role of heterogeneity is especially pronounced when policy is allowed to be targeted - directed at any individual bank in the distribution. We found that different policies generate different implications for the relative efficiency gains from targeting big vs small banks. Direct equity injections into large banks are more effective in stimulating aggregate output than injections into small banks. This is due to the shape of the MPL schedule which is increasing with bank equity capital. On the other hand, policies that relax the cost of capital or relative prices generally favor the small banks. This is due to the shape of the MPP and marginal costs schedules, which are both declining in bank size. Examples of such policies include direct lending and liquidity facilities as well as debt guarantee schemes.

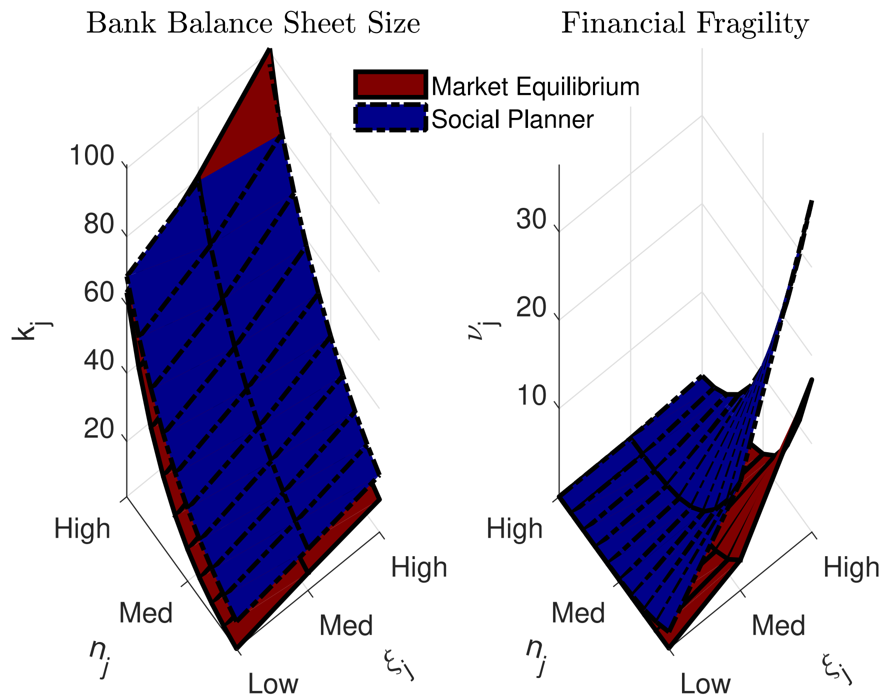
5 Optimal Micro-Prudential Policy

We now use our model to derive normative implications and quantify welfare losses in the market economy. We describe optimal allocations of a benevolent planner that faces the same market structure and environmental constraints but internalizes the impact of private, bank-level choices on aggregate returns. Specifically, the social planner cannot affect the elasticity θ or the market incompleteness parameter κ directly. The planner can, instead, improve on the market outcome by lowering equilibrium credit margins while pushing up the common component of the return on risky capital. Technically, we thus solve for the "constrained efficient" allocation in the language of Davila et al. (2012) and the general equilibrium literature.

5.1 Constrained Efficient Allocations

The dynamic problem of the social planner is identical to that of the monopolistic credit market with one exception. In this section, the planner picks the triple $\{k, d, p\}$ in order to maximize the

Figure 17: **Market Equilibrium and Constrained Efficient Allocations**



Notes: Market-based and social planner's allocations from the stationary equilibrium. The left panel shows the optimal choice of credit k(j) for the two cases. The right panel shows the equilibrium probability of bank default (insolvency), in annualized %. In both panels, the x-axis is bank net worth n(j) and y-axis is the realization of the idiosyncratic return process ξ(j) - the two bank-level state variables.

franchise value V while understanding that R^T is *endogenous* through the impact of the truple on R^k and P. Consider the law of motion of net worth that the social planner faces:

$$n_{t+1}(j) = R^T\Big(n_t(j), \xi_t(j), \{k_t(j), d_t(j), p_t(j)\}\Big)p_t(j)k_t(j) - \bar{R}_t(j)d_t(j) \tag{26}$$

Compare this formula to Equation 9 from the industry equilibrium. The difference is that R^T is no longer taken as given. Numerically, the banking problem is solved under the assumption that R^k and P are both polynomials in $\{n(j), \xi(j), \{k_t(j), d_t(j), p_t(j)\}\}$. We use projection methods to solve for the coefficients that are consistent with equilibrium. See Appendix D for more details on the numerical algorithm.[23]

Figure 17 presents key policy functions of the social planner in comparison with the market outcome. On the left panel, we plot the two-dimensional optimal choice of bank credit k(j). Comparing the two cases reveals that misallocation is present in the decentralized equilibrium

[23]Without loss of generality, we shut down the endogenous entry channel for simplicity. That is, the number of banks in the distribution is time-invariant.

along both the net worth and idiosyncratic risk dimensions. Specifically, the market outcome yields too *little* lending because of the aggregate credit supply externality. In addition, the wedge between the social optimum and the market outcome is larger for banks with low idiosyncratic returns realizations. This is driven by the pecuniary externality and market incompleteness.

The right panel of Figure 17 shows how equilibrium financial stability behaves in the two cases. We define (system-wide) financial stability as the average probability of bank default due to insolvency. Here we observe that the social planner's allocations induce a considerable increase in default risk. Low-net worth, high-$\xi(j)$ intermediaries become particularly more risky. This result is a case in point of the financial stability-competition trade-off (Hellman et al., 2000)[24]. Specifically, the social planner targets the distribution of credit margins but also reallocates equity capital away from smaller intermediaries that are fundamentally more prone to insolvency-triggering idiosyncratic shocks to begin with. Smaller banks with high realized idiosyncratic returns are more (market) levered in the constrained efficient equilibrium, and higher market leverage immediately increases the default likelihood.

5.2 Decentralization with Size- and Income-Dependent Taxation

We decentralize constrained efficient allocations with simple tax policies. Importantly, these policies are size-dependent because misallocation and credit margins correlate with the initial state of net worth. Theoretically, gross profit taxes are easy to work with because they target specifically the wedge in the bank-specific total portfolio return process and the law of motion of net worth. In addition, it is also typically less costly to obtain bank returns data from administrative sources than aggregate balance sheet quantity information. Specifically, we conjecture a size and idiosyncratic return specific tax rule $\tau\Big(n(j), \xi(j)\Big)$ and impose it in the decentralized equilibrium. Computationally, we assume that taxes are polynomials in n(j) and $\xi(j)$ and solve for coefficients that are consistent with a minimal distance between the equilibrium and the social planner allocations. See Appendix D for details. Note that negative taxes (subsidies) are allowed, which is important when working with underutilization of resources due to monopolistic competition. The law of motion of net worth with tax policies is now:

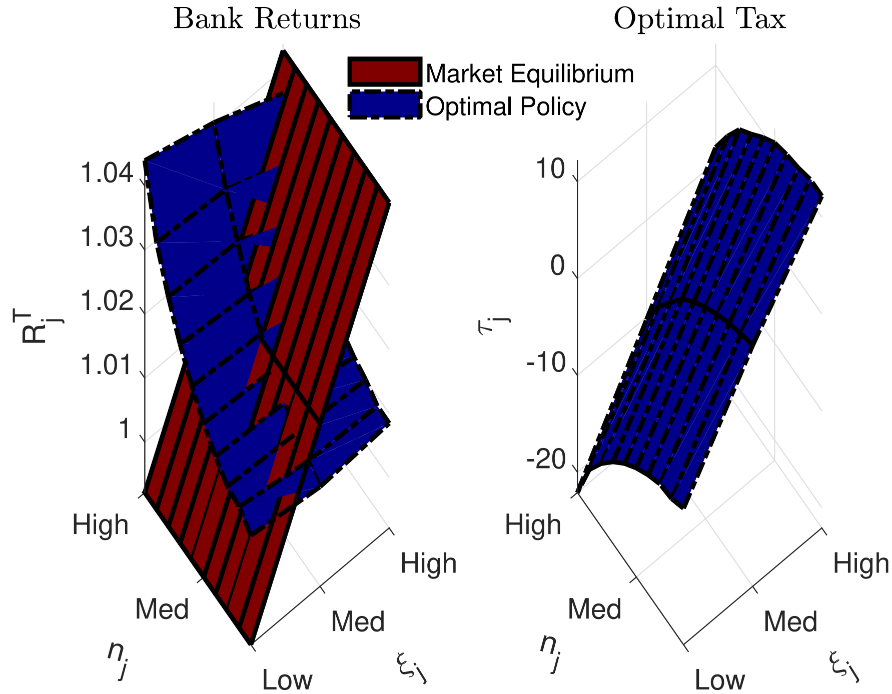
$$n_{t+1}(j) = R^T_t(j)\left[1 - \tau\Big(n(j), \xi(j)\Big)\right] p_t(j)k_t(j) - \bar{R}_t(j)d_t(j) \tag{27}$$

Effectively, on each point in the grid, we search for tax values that equalize socially optimal and market allocations.

Figure 18 presents the shape of the tax function. The left panel shows equilibrium bank returns

[24]Note that, in contrast, there is an alternative, competition-fragility view (Boyd and Nicolo, 2005; Ranciere et al., 2008)

Figure 18: **Optimal Micro-Prudential Policy**



Notes: Size-dependent bank policies that decentralize the social planner's constrained efficient allocations. The left panel shows equilibrium portfolio returns for the two cases. The right panel shows the optimal schedule of taxes on gross bank returns that - when applied to the market outcome - achieves constrained efficiency.

from the decentralized equilibrium and when optimal policy is implemented. The right panel shows the optimal size-dependent tax schedule that achieves the socially optimal ergodic distribution of portfolio returns. We label this tax instrument as an optimal *micro-prudential* policy. Many intermediaries in the state space receive a *subsidy*. The subsidy is the highest (in absolute terms) for *big* banks with *low* idiosyncratic returns. On the other hand, low-net worth intermediaries with high idiosyncratic returns face a positive tax. The intuition for this result is the following: marginal propensity to lend (MPL), which we characterized in the previous section, is highest precisely for the big banks with low returns. The social planner corrects equilibrium under-lending by stimulating credit supply of banks with the highest MPL. In general equilibrium, this maximizes aggregate output and household consumption. Normative policy implications are thus essentially in line with the targeted equity injections experiments which abstracted from any welfare considerations. In the stationary distribution, the annualized tax ranges from -27.49% to 21.01% with the average tax of about -5.04%, i.e. a subsidy.

Table 2 in the Appendix summarizes all the allocations in the decentralized and the constrained efficient equilibria. Most notably, welfare gains (measured in lifetime consumption units) from targeting credit markups are considerable and amount to 6.5%. Macroeconomic aggregates across the board are greater in the case of the social planner. Both the risk premium and the risk-free rate are higher. As mentioned above, the average default probability is substantially higher in the social planner's equilibrium. The net effect on consumer welfare is therefore not entirely obvious because our model lacks social, physical costs of default. To the extent that bank default could in fact be very costly in terms of real production units, the welfare gains from targeting credit market power or pecuniary externalities are potentially much smaller in practice. For example, Borio et al. (2012) estimate that financial crises can yield large, double-digit cumulative production-unit losses over a 3-year span. Coupled with a greater *frequency* of banking crises in the social planner's case, much of the steady-state welfare gain from optimal policy can get eroded quickly. Furthermore, we have assumed that decentralization of constrained efficient allocations is operationally costless. In reality, tax policy could be operationally costly on average with the cost potentially increasing in the size of the intermediary that is being regulated.

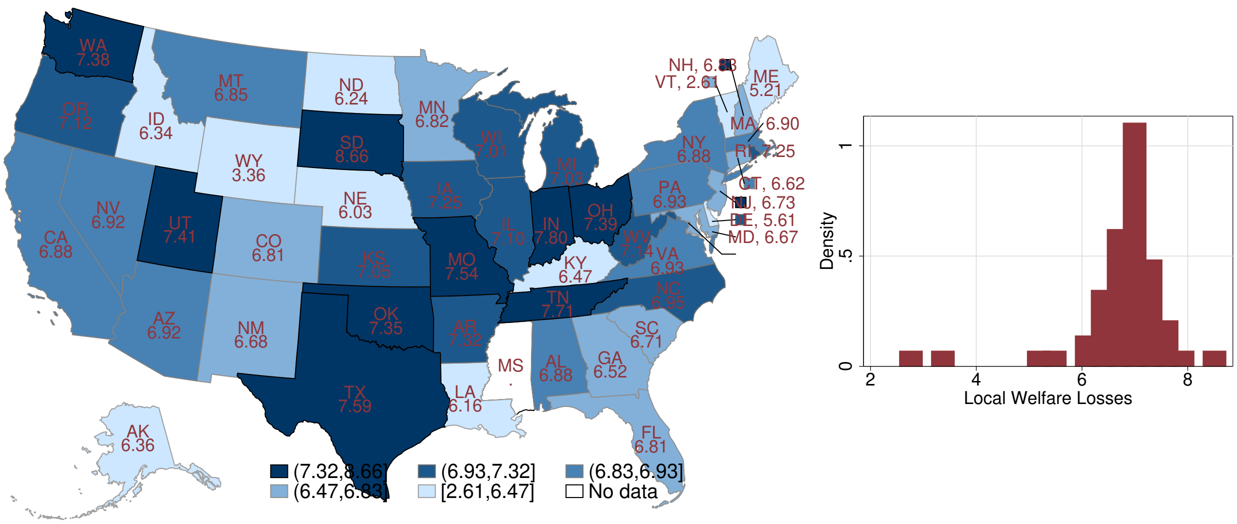
5.3 Welfare Losses from Local Credit Market Power

We conclude our normative analysis by computing the distribution of local, state-level welfare losses in the U.S. Jamilov (2020b) estimates reduced-form credit demand elasticities using branch-level interest rate data for U.S. banks in a quasi-experimental setting and an IV for local credit demand shocks. Linear regressions provide estimates of the average, nationwide credit demand elasticity θ. We used those estimates in the model calibration stage. However, that paper also relaxes the assumption that the elasticity is homogenous across U.S. states and runs 48 within-state, cross-branch regressions. This yields a distribution of 48 state-level demand elasticities. In this paper, we now feed this distribution into our model and compute the welfare cost from state-level, local values of θ_s, where index s captures an individual state.[25]

The distribution of absolute local welfare losses is shown in Figure 19. It plots results in the following way: the left panel presents a U.S. map, and the right panel - a one-dimensional histogram. For each state on the map, we report localized welfare losses in units of lifetime consumption conditional on the local state-specific θ_s. Regional variation in absolute welfare losses is substantial. In the extremes, numbers vary from 2.61% and 8.65%. The most affected states appear to be South Dakota, Indiana, Tennessee, and Texas. Again, we emphasize the competition-stability trade-off and that bank default is still assumed to be costless for consistency. In this particular exercise, if financial stability considerations are allowed to be state-specific, then variation of welfare losses could change. We leave this important issue for future research to follow.

[25] Technically, we solve the unregulated market and constrained efficient equilibria on a sparse grid of θ and use cubic spline interpolation.

Figure 19: **Welfare Losses from Local Credit Market Power**



Notes: welfare losses from local credit market power for 47 U.S. States. Credit demand elasticities are from Jamilov (2020b). Welfare implications are backed out based on the market and social planner's outcomes of the quantitative model. All numbers represent local welfare losses vs the constrained efficient counterfactual. For example, the nationwide average credit demand elasticity of 1.2 corresponds to the welfare loss that amounts to 6.5% of lifetime consumption.

This concludes the normative part of the paper. Optimal size- and income-dependent bank tax policy can decentralize constrained efficient allocations and improve consumer welfare considerably. Optimal micro-prudential policy yields qualitative predictions that are similar to the quantitative analysis of bank-level equity injections. With costly default, the financial competition-stability trade-off makes welfare gains from optimal policies potentially small.

6 Conclusion

In this article I show that bank heterogeneity has significant implications for the macroeconomy and the design of economic policies. I develop a novel quantitative, tractable, dynamic general equilibrium framework with endogenous bank heterogeneity to think about *targeted* bank regulation and supervision where policy can affect any *individual* bank in the stationary distribution. A central object in my analysis is the endogenous distribution of the *Marginal Propensity to Lend*: a sufficient statistic for the quantification of aggregate elasticities with respect to bank-level shocks. I characterize different types of credit policies, under various modelling scenarios and calibrations, and supplement the quantitative analysis with the analysis of *optimal* policy that decentralizes constrained efficient allocations. I find that with bank heterogeneity, different policy approaches have different cross-sectional and macroeconomic implications. Specifically, in terms of stimulating aggregate demand and production, equity injections directed at large intermediaries are more efficient. On the other hand, liquidity and direct lending facilities as well as deposit insurance schemes are more beneficial for small banks. I also find that optimal policy leads to similar qualitative conclusions as the targeted equity injection experiment.

My model is tractable and can be readily extended to include additional parts. First, an open-economy extension could be introduced, allowing us to study endogenous global financial cycles that are driven by heterogeneous intermediaries. Second, the model in its present form abstracts from aggregate risk. Jamilov and Monacelli (2020) build on my framework and introduce aggregate uncertainty and a dynamic distribution of bank size. They uncover novel channels of business cycle amplification that arise from dynamic bank heterogeneity. Finally, all heterogeneity in my model comes from marginal costs and idiosyncratic shocks as the credit markup is homogenous across agents and states of nature. An extension with credit markup heterogeneity could deliver an endogenous distribution of markups, potentially a time-varying object that is correlated with bank size and returns.

References

ADRIAN, T. AND H. S. SHIN (2010): "Liquidity and Leverage," *Journal of Financial Intermediation*, 19(3), 418–437.

——— (2014): "Procyclical Leverage and Value-at-Risk," *Review of Financial Studies*, 27(2), 373–403.

AIYAGARI, R. (1994): "Uninsured Idiosyncratic Risk and Aggregate Saving," *Quarterly Journal of Economics*, 109(3), 659–684.

ANDERSON, S., A. D. PALMA, AND J. THISSE (1989): "Demand for Differentiated Products, Discrete Choice Models, and the Characteristics Approach," *The Review of Economic Studies*, 56(1).

BEGENAU, J., S. BIGIO, J. MAJEROVITZ, AND M. VIEYRA (2020): "A Q-Theory of Banks," *Manuscript*.

BEGENAU, J. AND T. LANDVOIGT (2018): "Financial Regulation in a Quantitative Model of the Modern Banking System," *Working Paper*.

BENHABIB, B., A. BISIN, AND M. LUO (2018): "Wealth distribution and social mobility in the US: A quantitative approach," *American Economic Review*, Forthcoming.

BENIGNO, G., H. CHEN, C. OTROK, A. REBUCCI, AND E. R. YOUNG (2016): "Optimal capital controls and real exchange rate policies: A pecuniary externality perspective," *Journal of Monetary Economics*, 84.

BEWLEY, T. (1977): "The permanent income hypothesis: A theoretical formulation," *Journal of Economic Theory*, 16, 252 – 292.

BIANCHI, J. AND S. BIGIO (2018): "Banks, Liquidity Management and Monetary Policy," *Manuscript*.

BLANCHARD, O. AND N. KIYOTAKI (1987): "Monopolistic Competition and the Effects of Aggregate Demand," *American Economic Review*, 77(4).

BOCOLA, L. (2016): "The Pass-Through of Sovereign Risk," *Journal of Political Economy*, 124 (4).

BOISSAY, F., F. COLLARD, AND F. SMETS (2016): "Booms and Banking Crises," *Journal of Political Economy*, 124(2).

BORIO, C., E. KHRROUBI, C. UPPER, AND F. ZAMPOLLI (2012): "Systemic Banking Crises Database: An Update," *IMF Working Paper*, 12/163.

BOYD, J. AND G. D. NICOLO (2005): "The Theory of Bank Risk Taking and Competition Revisited," *Journal of Finance*, 60(3).

BRUNNERMEIER, M. AND L. PEDERSEN (2009): "Market Liquidity and Funding Liquidity," *Refiew of Financial Studies*, 22, 2201–2238.

BRUNNERMEIER, M. AND Y. SANNIKOV (2014): "A Macroeconomic Model with a Financial Sector," *American Economic Review*, 104(2), 379–421.

CABALLERO, R. AND G. LORENZONI (2014): "Persistent Appreciations and Overshooting: A Normative Analysis," *IMF Economic Review*, 62.

CAPELLE, D. (2019): "Competition vs. Stability: Oligopolistic Banking System with Run Risk," *Working Paper*.

CHRISTIANO, L. AND D. IKEDA (2013): "Leverage Restrictions in a Business Cycle Model," *NBER Working Paper 18688*.

COIMBRA, N. AND H. REY (2019): "Financial Cycles with Heterogeneous Intermediaries," *NBER Working Paper*, 23245.

CORBAE, D. AND P. D'ERASIMO (2019): "Capital Requirements in a Quantitative Model of Banking

Industry Dynamics," *NBER Working Paper*, 25424.

CORBAE, D. AND R. LEVINE (2018): "Competition, Stability, and Efficiency in Financial Markets," *Jackson Hole Symposium*.

CURDIA, V. AND M. WOODFORD (2010): "Credit Spreads and Monetary Policy," *Journal of Money, Credit and Banking*, 42.

——— (2016): "Credit Frictions and Optimal Monetary Policy," *Journal of Monetary Economics*, 84.

DAVILA, E. AND A. KORINEK (2017): "Pecuniary Externalities in Economies with Financial Frictions," *Review of Economic Studies*, 85, 352–395.

DAVILA, J., J. H. HONG, P. KRUSELL, AND J.-V. RIOS-RULL (2012): "Constrained Efficiency in the Neoclassical Growth Model With Uninsurable Idiosyncratic Shocks," *Econometrica*, 80.

DAVYDIUK, T. (2020): "Dynamic Bank Capital Requirements," *Journal of Financial Economics*, Forthcoming.

DIXIT, A. AND J. STIGLITZ (1977): "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, 67(3).

DRECHSLER, I., A. SAVOV, AND P. SCHNABL (2017): "The deposits channel of monetary policy," *Quarterly Journal of Economics*, 132, 1819–1876.

EGAN, M., A. HORTACSU, AND G. MATVOS (2017): "Deposit Competition and Financial Fragility: Evidence from the US Banking Sector," *American Economic Review*, 107(1).

FARHI, E. AND J. TIROLE (2017): "Shadow Banking and the Four Pillars of Traditional Financial Intermediation," *NBER Working Paper*, 23930.

FARHI, E. AND I. WERNING (2016): "A Theory of Macroprudential Policies in the Presence of Nominal Rigidities," *Econometrica*, 84(5).

GALAASEN, S., R. JAMILOV, R. JUELSRUD, AND H. REY (2020): "Granular Credit Risk," *NBER Working Paper 27994*.

GEANAKOPLOS, J. AND H. POLEMARCHAKIS (1985): "Existence, Regularity, and Constrained Suboptimality of Competitive Allocations When the Asset Market Is Incomplete," *Cowles Foundation Discussion Papers*, 764.

GERTLER, M. AND P. KARADI (2011): "A Model of Unconventional Monetary Policy," *Journal of Monetary Economics*, 58(1), 17–34.

GERTLER, M. AND N. KIYOTAKI (2010): "Financial Intermediation and Credit Policy in Business Cycle Analysis," *Handbook of Monetary Economics*, 3, 547–599.

GERTLER, M., N. KIYOTAKI, AND A. PRESTIPINO (2016): "Wholesale Banking and Bank Runs in Macroeconomic Modelling of Financial Crises," *Handbook of Macroeconomics*, 2.

——— (2020): "A Macroeconomic Model with Financial Panics," *Review of Economic Studies*, 87(1).

GOLDSTEIN, I., A. KOPYTOV, L. SHEN, AND H. XIANG (2020): "Bank Heterogeneity and Financial Stability," *NBER Working Paper 27376*.

GOPINATH, G., S. KALEMLI-OZCAN, L. KARABARBOUNIS, AND C. VILLEGAS-SANCHEZ (2017): "Capital Allocation and Productivity in South Europe," *Quarterly Journal of Economics*, 132, 1915–1967.

GORTON, G. AND A. METRICK (2010): "Regulating the Shadow Banking System," *Brookings Papers on Economic Activity*, Fall.

GREENWALD, B. AND J. STIGLITZ (1986): "Externalities in Economies with Imperfect Information and Incomplete Markets," *Quarterly Journal of Economics*, 101(2).

Hart, O. (1975): "On the optimality of equilibrium when the market structure is incomplete," *Journal of Economic Theory*, 11(3).

He, Z., B. Kelly, and A. Manela (2016): "Intermediary Asset Pricing: New Evidence from Many Asset Classes," *Journal of Financial Economics*, Forthcoming.

Hellman, T., K. Murdock, and J. Stiglitz (2000): "Liberalization, Moral Hazard in Banking, and Prudential Regulation: Are Capital Requirements Enough?" *American Economic Review*, 90(1).

Huggett, M. (1990): "The Risk-Free Rate in Heterogeneous Agent Economies," *Manuscript, University of Minnesota*.

Imrohoglu, A. (1996): "Costs of Business Cycles with Indivisibilities and Liquidity Constraints," *Journal of Political Economy*, 1364–83.

Jamilov, R. (2020a): "Bank Runs with Heterogeneous Intermediaries," *Working Paper*.

——— (2020b): "Credit Market Power: Branch-level Evidence from the Great Finanial Crisis," *Working Paper*.

Jamilov, R. and T. Monacelli (2020): "Bewley Banks," *CEPR Discussion Paper 15428*.

Jermann, U. and V. Quadrini (2013): "Macroeconomic Effects of Financial Shocks," *American Economic Review*, 102(1), 238–271.

Kaplan, G., B. Moll, and G. Violante (2018): "Monetary Policy According to HANK," *American Economic Review*, 108(3).

Kashyap, A. and J. Stein (2000): "What Do a Million Observations on Banks Say about the Transmission of Monetary Policy?" *American Economic Review*, 90(3).

Khwaja, A. and A. Mian (2008): "Tracing the Impact of Bank Liquidity Shocks: Evidence from an Emerging Market," *American Economic Review*, 98(4).

Korinek, A. and M. Nowak (2016): "Risk-Taking Dynamics and Financial Stability," *Manuscript*.

Korinek, A. and A. Simsek (2016): "Liquidity Trap and Excessive Leverage," *American Economic Review*, 106.

Krusell, P. and A. Smith (1998): "Income and Wealth Heterogeneity in the Macroeconomy," *Journal of Political Economy*, 106, 867–896.

Lorenzoni, G. (2008): "Inefficient Credit Booms," *Review of Economic Studies*, 75(3).

Maliar, L., S. Maliar, and F. Valli (2010): "Solving the incomplete markets model with aggregate uncertainty using the Krusellâ€“Smith algorithm," *Journal of Economic Dynamics and Control*, 34.

Martinez-Miera, D. and R. Repullo (2010): "Does Competition Reduce the Risk of Bank Failure?" *The Review of Financial Studies*, 23.

McFadden, D. (1984): "Econometric Analysis of Qualitative Response Models," *Grilliches, Z. and Intriligator, M. (eds) Handbook of Econometrics*, 2.

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

Nguyen, T. (2015): "Bank Capital Requirements: A Quantitative Analysis," *Working Paper*.

Nuno, G. and C. Thomas (2016): "Bank Leverage Cycles," *American Economic Journal: Macroeconomics*, Forthcoming.

Ranciere, R., A. Tornell, and F. Westermann (2008): "Systemic Crises and Growth," *Quarterly Journal of Economics*, 123(1).

Rios Rull, V., T. Takamura, and Y. Terajima (2020): "Banking Dynamics, Market Discipline and Capital Regulations," *Manuscript*.

SCHMITT-GROHE, S. AND M. URIBE (2003): "Closing small open economy models," *Journal of International Economics*, 61.

SHLEIFER, A. AND R. VISHNY (1988): "The Efficiency of Investment in the Presence of Aggregate Demand Spillovers," *Journal of Political Economy*, 96(6).

STAVRAKEVA, V. (2019): "Monetary Policy Spill-Overs in Emerging Markets and Imperfect Banking Sector Competition," *Manuscript*.

——— (2020): "Optimal Bank Regulation and Fiscal Capacity," *Review of Economic Studies*, 87(2).

TAUCHEN, G. (1986): "Finite state markov-chain approximations to univariate and vector autoregressions," *Economics Letters*, 20.

WHEELOCK, D. (2011): "Banking Industry Consolidation and Market Structure: Impact of the Financial Crisis and Recession," *Federal Reserve Bank of St. Louis Review*, 93(6).

A Proofs

A.1 Proposition 1 (Scale Variance)

Guess that the solution to the dynamic problem is a value function $V\Big(n(j), \xi(j)\Big) = \vartheta\Big(n(j), \xi(j)\Big)n(j)$. Define the default risk-adjusted stochastic discount factor $\tilde{\Lambda} = \Big[\Big(1 - \upsilon(j)\Big)\Lambda\Big(1 - \sigma + \sigma\vartheta\Big(n'(j), \xi'(j)\Big)\Big)\Big]$. The solution to the program is a system of equations:

$$\mathbb{E}\left[\tilde{\Lambda}'\left(R^{T'}(j) - \bar{R}(j)\right)\right] = \lambda\varphi\Big(n(j), \xi(j)\Big)$$

$$\varphi\Big(n(j), \xi(j)\Big)\left[\vartheta\Big(n(j), \xi(j)\Big) - \lambda\phi(j)\right] = 0$$

Substituting the optimality conditions together with the guess into the objective function gives

$$\vartheta\Big(n(j), \xi(j)\Big) = \varphi\Big(n(j), \xi(j)\Big)\vartheta\Big(n(j), \xi(j)\Big) + \mathbb{E}\Big(\tilde{\Lambda}'\Big)(\bar{R}(j) - \frac{\frac{1}{\zeta_1}k(j)^{\zeta_2}}{n(j)})$$

Solving for $\vartheta\Big(n(j), \xi(j)\Big)$ yields

$$\vartheta\Big(n(j), \xi(j)\Big) = \frac{\mathbb{E}\Big(\tilde{\Lambda}'\Big)\left(\bar{R}(j) - \frac{\frac{1}{\zeta_1}k(j)^{\zeta_2}}{n(j)}\right)}{1 - \varphi\Big(n(j), \xi(j)\Big)}$$

And the Lagrange multiplier on the leverage constraint is

$$\varphi\Big(n(j), \xi(j)\Big) = \max\left[1 - \frac{\mathbb{E}\Big(\tilde{\Lambda}\Big)\left(\bar{R}(j) - \frac{\frac{1}{\zeta_1}k(j)^{\zeta_2}}{n(j)}\right)}{\lambda\phi(j)}, 0\right]$$

The result follows from (a) the fact that market leverage is $\phi(j) = k(j)^{\frac{\theta-1}{\theta}}K^{\frac{1}{\theta}}Pn(j)^{-1}$ (b) and the previously defined augmented stochastic discount factor $\tilde{\Lambda}$. The guess is verified if $\varphi\Big(n(j), \xi(j)\Big) < 1$. Net worth-dependency is guaranteed by $\zeta_1 \neq 0$,$\zeta_2 \neq 1$ so that each bank with different n(j) chooses a different leverage ratio $\phi(j)$. Furthermore, with $\kappa > 0$, $\phi(j)$ also explicitly depends on $\xi(j)$. Note that θ does not impact scale-dependency but does change the level and curvature of the $\vartheta\Big(n(j), \xi(j)\Big)$ surface. □

A.2 Proposition 2 (Aggregate Credit Supply Externality)

Assumptions: Bank-level choices are made while $\bar{R}(j)$, $R^T(j)$, $\nu(j)$ are taken as given. Leverage constraint is slack. Without loss of generality, assume $\zeta_1 = \zeta_2$.

First, prove that the bank price-setting rule is:

$$\frac{p(j)}{P} = \left[\frac{\theta}{\theta-1}\frac{1}{\left(1-\nu(j)\right)R^T(j)-\bar{R}(j)}\frac{1}{P}K^{\zeta_2-1}\right]^{\frac{1}{1+\theta(\zeta_2-1)}}$$

Each bank j maximizes its franchise value:

$$\max_{k(j)} \left(1-\upsilon(j)\right)R^T(j)p(j)k(j) - \bar{R}\left(p(j)k(j)-n(j)\right) - \frac{1}{\zeta_2}k(j)^{\zeta_2} - FC \quad \text{s.t.} \quad k(j) = \left(\frac{p(j)}{P}\right)^{-\theta}K$$

The first order condition is

$$\left[\left(1-\nu(j)\right)R^T(j)-\bar{R}(j)\right]\left[p(j)+k(j)\frac{\partial p(j)}{\partial k(j)}\right] = k(j)^{\zeta_2-1}$$

Assume that the impact of p(j) on the aggregate index P is not internalized. The elasticity is:

$$\frac{\partial k(j)}{\partial p(j)}\frac{p(j)}{k(j)} = -\theta$$

The price-setting rule given marginal costs is

$$p(j) = \frac{\theta}{\theta-1}MC(j)$$

where $\frac{\theta}{\theta-1}$ is the constant markup over the (endogenous) marginal cost MC(j), given by:

$$MC(j) := \frac{1}{\left(1-\nu(j)\right)R^T(j)-\bar{R}(j)}\left[\left(\frac{p(j)}{P}\right)^{-\theta}K\right]^{\zeta_2-1}$$

Substituting MC(j) into the price-setting rule yields:

$$\frac{p(j)}{P} = \left[\underbrace{\frac{\theta}{\theta-1}}_{\text{Markup}}\underbrace{\frac{1}{\left(1-\nu(j)\right)R^T(j)-\bar{R}(j)}\frac{1}{P}K^{\zeta_2-1}}_{\text{Marginal Cost}}\right]^{\frac{1}{1+\theta(\zeta_2-1)}}$$

In a symmetric equilibrium, prove that the aggregate rate-setting rule is:

$$\bar{R} = \left(1-\nu\right)R^T - \frac{\theta}{\theta-1}\frac{1}{P}K^{\zeta_2-1}$$

The result is achieved trivially if we impose symmetry into the result of Part 1. This establishes the downward-sloped relationship between aggregate demand K and the price of deposits $\bar{R}$.

The upward sloped supply of funds curve is the FOC of the household wrt deposit holdings under the symmetry assumption.

$$\bar{R} = \frac{1 - \Lambda \upsilon x R^T}{\left(1 - \upsilon\right)\Lambda}$$

Equilibrium exists for a given $\zeta_2 > 1$. Decrease in θ shifts the demand schedule leftwards. Everything else equal, for any finite $\theta > 1$, equilibrium demand (K) is strictly less than in the case of perfect competition or an infinite θ. □

A.3 Asset Pricing Implications

The banker's Euler equation can be re-formulated into a classical asset pricing formula for the risk premium

$$\mathbb{E}_t\left[R^T_{t+1}(j) - \bar{R}_t(j)\right] = \frac{\lambda\varphi(j)}{\mathbb{E}_t\left(\hat{\Lambda}_{t+1}(j)\right)} + \text{cov}\left[-\frac{\hat{\Lambda}_{t+1}(j)}{\mathbb{E}_t\left(\hat{\Lambda}_{t+1}(j)\right)}, \left(1 - \upsilon(j)\right)R^T_{t+1}(j)\right] \quad \forall j$$

Where $\varphi(j)$ is the Lagrange multiplier on the moral hazard (leverage) constraint. Note that the equation must hold for every bank (j) in the distribution. Excess returns in the model arise for two reasons. First, if the hard leverage constraint binds for any given variety j, then external funds are harder to obtain. This is the canonical liquidity-induced external finance premium. Second, fluctuations in default risk are generally positively correlated with the SDF. High default risk implies low expected returns on investment. This negatively impacts equilibrium consumption and thus requires ex-ante compensation.

We now explore the impact of credit market power (CMP) on average risk premia in equilibrium. First, greater CMP (lower θ) provides more incentives for the bank to raise prices p(j). In the aggregate, prices go up and investment demand goes down for all discount rate levels, which is the aggregate credit supply externality from Section 2.11. Lower demand, in turn, shifts the distribution of bank size to the left - the average bank is smaller. This makes it more likely that the leverage constraint binds, because he constraint restricts *market* leverage. As a result, liquidity premia go up. At the same time, bank default risk goes down because of diminished appetite for risk-taking and the competition-stability trade-off. This lowers the default risk premium. Quantitatively, it turns out that liquidity premia fluctuations are rather small (less than 10% of the total premium), because the leverage constraint is rarely binding except for the very small banks whose distributional weight is not big enough to impact the aggregate. As a result, higher CMP causes the total risk premium to fall.

A.4 Size-Leverage Complementarity

Recent contributions by Gopinath et al. (2017), Coimbra and Rey (2019), and Jamilov and Monacelli (2020) provide empirical evidence for a positive cross-sectional correlation between firm size and (book) leverage. In this paper, book leverage indeed grows with bank size. Market leverage, however, falls with size because big banks are able to set lower relative prices. Define leverage as ratio of book assets to book net worth. Abstract from non-interest and fixed costs. We can show that if the leverage constraint binds then leverage scales with bank size as summarized by the following proposition

Proposition 3. *If (a) $\theta > 1$ and (b) $\lambda > \mathbb{E}_t\left[\tilde{\Lambda}_{t+1}(j)\left(R^T_{t+1}(j) - \bar{R}(j)\right)\right]$ then*

$$\frac{\partial\phi(j)}{\partial n(j)} > 0$$

Proof. Assuming the leverage constraint binds, solve for book leverage

$$\phi(j) = \frac{\nu_a(j)}{\lambda - \mu_a(j)}$$

where $\mu_a(j)$ and $\nu_a(j)$ are defined in the dynamic problem 2. Simplify, taking the credit demand function as given

$$\phi(j) = n(j)^{\frac{1}{\theta-1}}\left[\frac{\tilde{\Lambda}(j)\bar{R}(j)}{K_t^{1/\theta}P_t\left(\lambda - \tilde{\Lambda}_{t+1}(j)\left(R^T_{t+1}(j) - \bar{R}(j)\right)\right)}\right]^{\frac{\theta}{\theta-1}}$$

Leverage-size dependency in partial equilibrium can be summarized as

$$\frac{\partial\phi(j)}{\partial n(j)} = \frac{1}{\theta - 1}\mathbf{A}n(j)^{-\frac{\theta-2}{\theta-1}}$$

where $\mathbf{A}$ is a collection of equilibrium terms which are taken as given during optimization:

$$\mathbf{A} = \left[\frac{\tilde{\Lambda}(j)\bar{R}(j)}{K_t^{1/\theta}P_t\left(\lambda - \tilde{\Lambda}_{t+1}(j)\left(R^T_{t+1}(j) - \bar{R}(j)\right)\right)}\right]^{\frac{\theta}{\theta-1}}$$

As long as the two conditions a) $\theta > 1$ and b) $\lambda > \tilde{\Lambda}_{t+1}(j)\left(R^T_{t+1}(j) - \bar{R}(j)\right)$ are satisfied, can show that $\mathbf{A} > 0$ and $\frac{\partial\phi(j)}{\partial n(j)} > 0$. The first condition is satisfied trivially, and the second one is always satisfied quantitatively if we set λ to any value above 0.053 given our baseline calibration. In addition, if $\theta > 2$ then $\frac{\partial^2\phi(j)}{\partial n(j)^2} < 0$, i.e. a positive and concave relationship between bank size and leverage is obtained, which matches the general patterns observed empirically.

□

A.5 Size-Dependent Responses to Cost of Funds Shocks

In a seminal empirical contribution, Kashyap and Stein (2000) show that small banks are more responsive to monetary policy shocks. We can verify that the same pattern of heterogeneous responses to cost of funds shocks holds in our model. Suppose monetary policy can directly influence the equilibrium cost of funds, i.e. the interest rate on deposits. Define leverage as the ratio of book assets to book net worth. The following proposition summarizes the result

Proposition 4. *If* $\mathbb{E}_t\left[\tilde{\Lambda}_{t+1}(j)\left(R^T_{t+1}(j)-\bar{R}(j)\right)\right] > \left[\lambda-\bar{R}(j)\right]$ *then*

$$\frac{\partial\phi(j)}{\partial n(j)\partial\bar{R}(j)} < 0$$

Proof. Following Proposition 3, size-leverage dependency is given by

$$\frac{\partial\phi(j)}{\partial n(j)} = \frac{1}{\theta-1}\left[\frac{\tilde{\Lambda}_{t+1}(j)\bar{R}(j)}{K_t^{1/\theta}P_t\left(\lambda-\tilde{\Lambda}(j)\left(R^T_{t+1}(j)-\bar{R}(j)\right)\right)}\right]^{\frac{\theta}{\theta-1}} n(j)^{-\frac{\theta-2}{\theta-1}}$$

In partial equilibrium, size-dependent sensitivity to cost of funds shocks $\bar{R}(j)$ is summarized by

$$\frac{\partial\phi(j)}{\partial n(j)\partial\bar{R}(j)} = \frac{\partial\left[\frac{1}{\theta-1}n(j)^{-\frac{\theta-2}{\theta-1}}\left[\frac{\tilde{\Lambda}\bar{R}(j)}{K_t^{1/\theta}P_t\left(\lambda-\tilde{\Lambda}_{t+1}(j)\left(R^T_{t+1}(j)-\bar{R}(j)\right)\right)}\right]^{\frac{\theta}{\theta-1}}\right]}{\partial\bar{R}(j)}$$

Can show that $\frac{\partial\phi(j)}{\partial n(j)\partial\bar{R}(j)} < 0$ as long as $\Lambda\left(R^T_{t+1}(j)-\bar{R}(j)\right) > \left[\lambda-\bar{R}(j)\right]$. This condition is always satisfied in the stationary equilibrium as long as expected excess returns are not <0 for a participating banker (which is never the case by construction, otherwise the banker doesn't enter) and $\lambda < 1$.

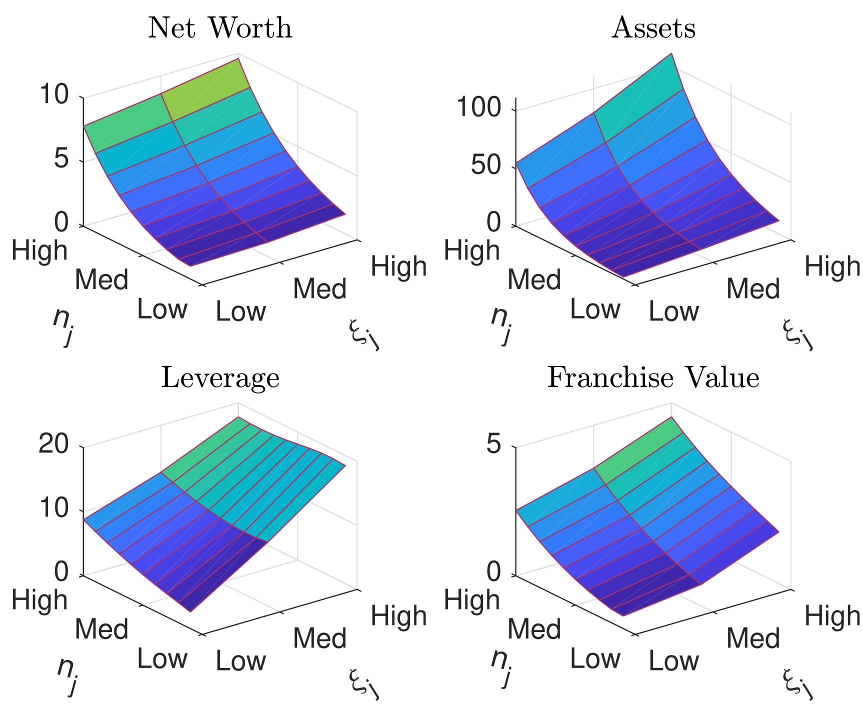
□

B Additional Model Results

B.1 Policy Functions

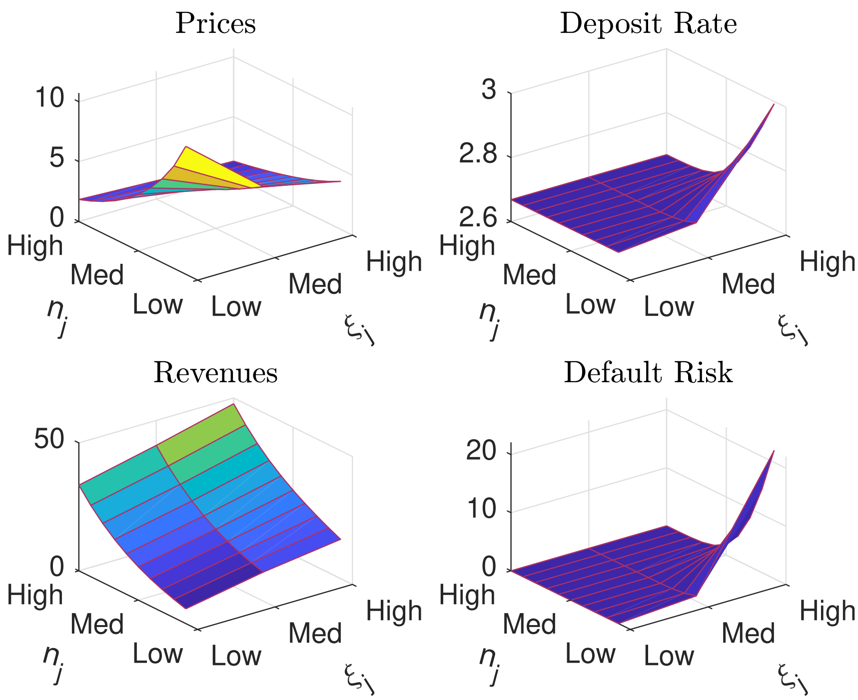
Figure 20 visualizes the relevant policy functions for quantities. Diagrams are two-dimensional since the state vector includes bank net worth n(j) and idiosyncratic rate of return ξ(j). Bank net worth growth, total assets, and franchise values V(j) all increase in both size and realized returns. (Book) leverage is increasing in bank size, which is consistent with existing empirical evidence.

Figure 20: **Model Policy Functions - Quantities**



Notes: This figure plots equilibrium model policy functions for quantities and book leverage. Bank size is defined as net worth n(j). Leverage is defined as the ratio of book assets over book equity $\phi(\mathrm{j}) = \frac{\mathrm{k(j)}}{\mathrm{n(j)}}$. The franchise value corresponds to V(j) in the paper. All plots are two-dimensional: the x and y axes represent net worth and realized idiosyncratic returns ξ(j), respectively.

Figure 21: **Model Policy Functions - Prices**



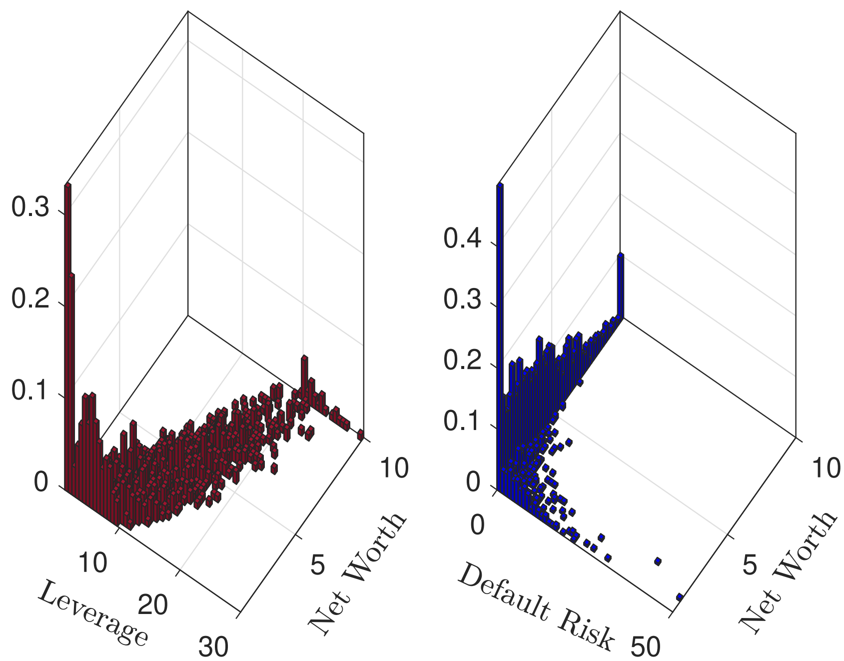
Notes: This figure plots equilibrium model policy functions for quantities and book leverage. Funding cost is the interest rate on deposits $\bar{R}(j)$. Prices in level are p(j). Revenues are defined as the product of prices and quantities p(j)k(j). Default risk corresponds to the probability of insolvency $\nu(j)$, in percent. All plots are two-dimensional: the x and y axes represent net worth and realized idiosyncratic returns $\xi(j)$, respectively.

Figure 21 plots the size-dependent price-setting rules and proxies of default risk. Prices in level (j) are decreasing and concave in bank size, which is consistent with the empirical evidence that loan margins are decreasing in bank size (Jamilov and Monacelli, 2020). Larger banks are also more profitable, which is intuitive. Probability of default due to insolvency $\upsilon(j)$ is the highest for small banks with low levels of net worth; it can reach levels as high as 20 percent annualized.

B.2 Stationary Cross-Sectional Distributions

We now visualize the ergodic cross-sectional distributions that are associated with the stationary industry equilibrium. We zoom in first on the cross-section of risk-taking in Figure 22. We notice a positively skewed distribution of bank leverage, which is consistent with the data (Coimbra and Rey, 2019; Jamilov and Monacelli, 2020). It's centered around 11 - the steady state target for the average - but is very disperse. In particular, some larger banks choose leverage ratios in excess of 15 or even 25, consistent with the data on the pre-crisis build-up of financial leverage (Gorton and Metrick, 2010). Bank default risk is concentrated, as seen from policy functions, around 0 but is falling in bank size. Some very small institutions are in considerable distress as likelihood of default reaches 20+ percent annualized.

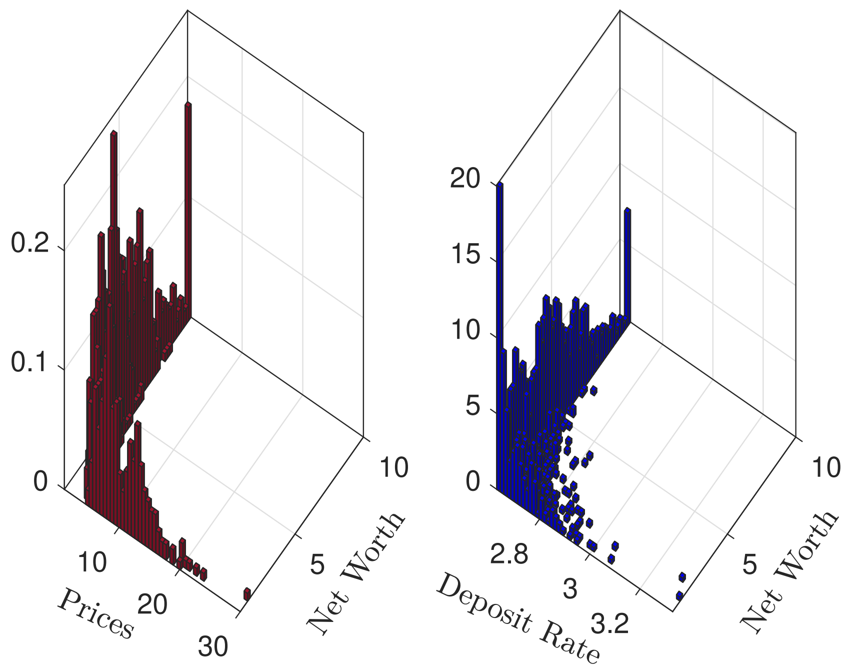
Figure 22: **Cross-Sectional Distributions - Quantities**



Notes: These figures present cross-sectional distributions of key variables in the structural model. All distributions are plotted on two-dimensional histograms where the y axis always represents bank size (net worth) and x axes represent variables that are labeled correspondingly. Figure plots book leverage, defined as the ratio of book assets over book equity $\phi(j) = \frac{k(j)}{n(j)}$ and probability of insolvency $\nu(j)$ - default risk. Figures plot the pdf normalization.

Figure 23 shows the distributions of bank-level prices and deposit rates. Prices are centered around roughly 1.436 and decline with bank size. The funding rate distribution is a mirror image of the distribution of default risk. That is, interest rates on deposits are high if a bank is at greater insolvency risk. The distribution is thus centered around the steady-state risk-free level but spikes up for smaller institutions which must pay a premium for being more risky. This premium can be as high as 10 basis points per quarter.

Figure 23: **Cross-Sectional Distributions - Prices**



Notes: These figures present cross-sectional distributions of key variables in the structural model. All distributions are plotted on two-dimensional histograms where the y axis always represents bank size (net worth) and x axes represent variables that are labeled correspondingly. Figure plots bank-level prices p(j) and interest rates on deposits $\hat{R}(j)$ - the funding rate. Figures plot the pdf normalization.

Table 2: **Comparing Equilibrium and Socially Optimal Allocations**

	Market Equilibrium	Optimal Policy
Welfare	1.0000	1.0649
Consumption	2.1405	2.2794
Output	2.4375	2.7499
Investment	0.2970	0.4705
Bank Leverage	11.1124	17.1849
Risk Premium	5.0538	8.3487
Risk-Free Rate	2.6024	2.7795
Credit Margin	1.4361	1.0009
Average Default Probability	1.2147%	1.8852%

Notes: This table compares allocations of the social planner that are decentralized with size-dependent tax policies applied to the unregulated market outcome. Details of the social planner problem and of the optimal tax function are in Section 5. All probabilities are annualized percent. Welfare is in terms of relative permanent consumption units.

C Microfounding CES Monopolistic Financial Intermediation

This section provides a brief theoretical foundation for the representative-agent capital goods producer's CES credit demand system. My approach follows closely McFadden (1984) and Anderson et al. (1989). Assume there are M borrowers and H banks. Each banker i posts its price schedule. Each borrower j observes the price menu and receives an idiosyncratic preference shock ϵ_{ij} which is borrower-creditor specific.

Assume the production function of a borrowing firm j is $\log k(j)$. All borrowers are indexed by their favorite bank branch $\bar{\epsilon}$. They suffer disutility measured in Euclidean distance between their preferred type and any given type i. Unit cost of that disutility, as well as the distance between varieties have been set to unity. Profit function of each firm takes on the following form.

$$Q_i(\bar{\epsilon}; k_i) = \underbrace{\log k_i + Y - p_i k_i}_{\text{Homogenous across j}} - \underbrace{\sum_{k=1}^{M} \left(\bar{\epsilon}^k - \epsilon_i^k\right)^2}_{\text{Heterogeneous across j}} \qquad i=1\ldots H \tag{28}$$

The first term in the equation is common across all borrowers and is bank-specific. The second term is the bank-borrower fixed effect that captures disutility from not borrowing from the ideal branch $\bar{\epsilon}$. Without loss of generality, we impose $M = H - 1$ for analytical convenience. We define *credit market access* as the set of consumers that are indifferent between borrowing from any two branches n:

$$\bar{E}^j = \frac{\log \frac{p(j)}{p_n}}{4} \tag{29}$$

The choice variables are (a) which branch to borrow from and (b) how much k(j). The price of the loan p(j) corresponds to the price on a claim on returns to capital in the main text. Y is endogenous real income that in equilibrium will equal K, i.e. the book value of capital after assembly and aggregation.

Every borrower in the credit market access space borrows $\frac{1}{p_n}$ units of differentiated loans from bank n. Demand k_n becomes:

$$k_n = \frac{1}{p_n} \int_{-\infty}^{\bar{\epsilon}_1} \ldots \int_{-\infty}^{\bar{\epsilon}_{n-1}} f\left(\bar{E}^j\right) d\bar{\epsilon} \tag{30}$$

Where we assume that k_n is strictly positive for all prices p(j), is n-1 times continuously differentiable, and all cross-price derivatives are positive for all i and j. Solution for the credit demand function above involves taking n-1 derivatives of k_n w.r.t. $p_1, \ldots, p_{H-1}$:

$$\frac{\partial^{H-1} k_n}{\partial p_1 \ldots \partial p_{H-1}} = \frac{1}{p_1 \ldots p_n} 4^{1-H} f\left(\bar{E}^j\right) \tag{31}$$

We assume that the firm borrower demand function is logistic in the cross-price differential $p(j) - p(i)$ for any two branches i and j. The density function associated with a logit credit demand is given by:

$$f(\bar{\epsilon}) = H\frac{4}{\bar{\theta}}^{H-1}(H-1)!\,\frac{\prod_{i=1}^{H-1}\exp\left(-4/\bar{\theta}\epsilon^{i}\right)}{\left[1+\sum(j)^{H-1}\exp\left(-4/\bar{\theta}\epsilon^{j}\right)\right]^{H}} \tag{32}$$

Plugging our model-specific credit market access variable into the logit density, and evaluating the first order condition yields

$$\frac{\partial^{H-1}k_n}{\partial p_1\ldots\partial p_{H-1}} = H\bar{\theta}^{1-H}(H-1)!\,\frac{\prod(j)^{H}p(j)^{-1/\bar{\theta}-1}}{\left[\sum(j)^{H}p(j)^{-1/\bar{\theta}}\right]^{H}} \tag{33}$$

Integrating gives us optimal credit demand

$$k_n = Hp_n^{-1/\bar{\theta}-1}\left[\sum(j)^{H}p(j)^{-1/\bar{\theta}}\right]^{-1} \tag{34}$$

Now, we impose the following parameter restriction: $\bar{\theta} = \frac{1}{\theta-1}$. Furthermore, impose the accounting identity that the total sum of firm-level loans is equal to the income of the representative capital goods producer: $Hk(j) = K$. We retrieve the CES credit demand function of firm j in main text:

$$k(j) = \left(\frac{p(j)}{P}\right)^{-\theta}K \tag{35}$$

We have thus shown that the representative-agent capital goods producer setup in main text is isomorphic to a heterogeneous-borrower environment with idiosyncratic preferences for branch amenities. The logit parameter $\bar{\theta}$ captures the variance of borrower preferences and maps conveniently to the CES elasticity θ. The relationship is inversed, so a higher $\bar{\theta}$ is associated with a lower elasticity of credit supply, i.e. greater credit market power. In the limit, if $\bar{\theta} \rightarrow \infty$ we recover a case with a single pure monopoly provider of credit. As $\bar{\theta} \rightarrow 0$ we recover the case of perfect competition in the banking sector. Because the problem discussed in this section is static, and assuming the distribution of shocks is time-invariant, heterogeneous firms would solve the same static problem every period and arrive at the same solution. It's therefore convenient, as we do in the main text, to work the representative-agent representation of this distribution.

D Solution Algorithm

In this section we lay out the numerical algorithm that is used to solve different variants of the model. Section D.1 describes how to solve the baseline unregulated market economy. Section D.3 solves the constrained efficient allocations of the social planner. Finally, D.3 solves the regulated economy which decentralizes the social planner's solution.

D.1 Unregulated Market Equilibrium

There are four basic computational challenges when solving the model in Section 2. First, we must solve the dynamic optimization problems of the financial intermediary and of the household. Because the banking sector is not scale invariant, individual bank characteristics matter for aggregation. The individual state vector is $\{n(j), \xi(j)\}$. We use value function iteration to solve the banking problem. For the household's problem we use time iteration on the household's Euler equation for deposit holdings.

Second, banks face an occasionally binding constraint on leverage. We deal with this issue the following way. On each point of the idiosyncratic state space we solve for optimal balance sheet choices while assuming the constraint binds. Then, we back out the Lagrange multiplier on the constraint. If the constraint is, in fact slack, we solve the problem again using numerical global maximization routine. If it is binding, then we continue the program.

Third, the market for deposit holdings must clear between bankers and the household. Because there is no deposit insurance, households price the full distribution of bank default risk into the menu of deposit rates. We iterate on the equilibrium deposit rates in the outer loop of the program. In each iteration, the newly constructed household's stochastic discount factor Λ - an endogenous aggregate state - is adopted by the banks.

Finally, there are 2 key aggregate (static) variables that are needed to pin down R^k: K and P. We use a variant of the stochastic simulation approach as in Maliar et al. (2010) to pin them down. Specifically, after solving for both the household's and banker's problems, we run a long simulation with N=1 intermediary and T=20,000 periods while using the newly computed policy functions of the incumbent banker. The distribution grants us new measures of aggregate demand of the incumbent. We combine these with demand from new entrants, which gives us new candidates for aggregate capital and prices. We require convergence tolerances of 10^{-6} for general equilibrium deposit rates, and 10^{-5} for the bankers' and household value functions.[26] Below we list the state variables of the model and sketch the solution algorithm.

Exogenous idiosyncratic shocks: $\{\xi(j)\}$. Exogenous idiosyncratic states: $\{n(j)\}$

Endogenous idiosyncratic states: $\{\upsilon(j), \bar{R}(j)\}$. Endogenous aggregate states: $\{K, P, \Lambda\}$

Algorithm - Decentralized Equilibrium

1. Guess some initial values for aggregate endogenous states $\{K, P, \Lambda\}$. Compute R^k. Guess some initial values for idiosyncratic endogenous states $\{\upsilon(j), \bar{R}(j)\}$

[26]Importantly, there is no aggregate risk in the model.[27] We therefore do not need to track a dynamic distribution of bank net worth.

2. Solve the financial intermediation problem
   (a) Use value function iteration. On each grid point, assume the leverage constraint binds.
   (b) Construct the Lagrange multiplier. If constraint indeed binds, proceed.
   (c) If constraint is slack, solve the problem again using a numerical minimization routine
3. Simulate the problem of the incumbent. Run a simulation of N=1 bankers and T=20,000 periods.
4. Solve the new entry problem. Determine the mass of entrants and their aggregate demand for capital.
5. Compute economywide new guesses for K' and P'. Construct a new $R^{k'}$.
6. Calculate the probability of bank default on each grid point using the newly computed policy functions and distributional aggregates. This gives new $\{\upsilon'(j), \bar{R}'(j)\}$
7. Solve the household's problem. Get new Λ'.
8. Compare $\{\bar{R}(j)\}$ with $\{\bar{R}'(j)\}$. If maximal error cross all grid points is within the tolerance level, general equilibrium is solved. If not, update $\{\bar{R}(j)\}$. Return to Step 2 and continue the iteration.

D.2 Constrained Efficient Equilibrium

In order to solve for constrained efficient (socially optimal) allocations, we must make one adjustment to the algorithm. The only difference between the decentralized solution and the social planner is that the latter internalizes the impact of private choices on aggregate returns. We operationalize this using projection methods. Specifically, we assume that both K and P are polynomials in n(j), ξ(j), and the choice of k(j). That is:

$$\begin{aligned} K &= \alpha_0^k + \alpha_1^k n(j) + \alpha_2^k \xi(j) + \alpha_3^k k\Big(n(j), \xi(j)\Big) \\ P &= \alpha_0^p + \alpha_1^p n(j) + \alpha_2^p \xi(j) + \alpha_3^p k\Big(n(j), \xi(j)\Big) \end{aligned}$$

The objective of the projection is then to find the optimal vector of coefficients $\{\alpha^k, \alpha^p\}$. We now describe the steps of the algorithm below.

Algorithm - Constrained Efficient Equilibrium

1. Guess some initial values for $\{\alpha^k, \alpha^p\}$.
2. Guess some initial values for aggregate endogenous states $\{K, P, \Lambda\}$. Compute R^k. Guess some initial values for idiosyncratic endogenous states $\{\upsilon(j), \bar{R}(j)\}$. Decentralized equilibrium solution works as a good first guess
3. Solve the financial intermediation problem under the social planner

(a) Given $\{\alpha^k, \alpha^p\}$, treat R^k as endogenous to the states and to the candidate choices of k(j). Use a numerical minimization routine to solve for the optimal k(j) on each grid point.

(b) On each grid point, first assume the leverage constraint binds.

(c) Construct the Lagrange multiplier. If constraint indeed binds, proceed.

(d) If constraint is slack, solve the problem again using a numerical minimization routine. Keep treating R^k as endogenous to states and choices

4. Simulate the problem of the incumbent. Run a simulation of N=1 bankers and T=20,000 periods. Run a linear regression of capital holdings k(j) on a constant, lagged net worth $n_{t-1}(j)$, lagged $\xi_{t-1}(j)$, and lagged capital holding $k_{t-1}(j)$. Do the same for the p(j). Compute new guesses for $\{\alpha^{k'}, \alpha^{p'}\}$.

5. Solve the new entry problem. Determine the mass of entrants and their aggregate demand for capital.

6. Compute economywide new guesses for K' and P'. Construct a new $R^{k'}$.

7. Calculate the probability of bank default on each grid point using the newly computed policy functions and distributional aggregates. This gives new $\{\upsilon'(j), \bar{R}'(j)\}$

8. Solve the household's problem. Get new Λ'.

9. Compare $\{\alpha^k, \alpha^p\}$ with $\{\alpha^{k'}, \alpha^{p'}\}$. And compare $\{\bar{R}(j)\}$ with $\{\bar{R}'(j)\}$. If the maximal errors across all grid points are within the tolerance level, constrained efficient equilibrium is solved. If not, update $\{\alpha^k, \alpha^p\}$ and $\{\bar{R}(j)\}$. Return to Step 3 and continue the iteration.

D.3 Decentralization with Size-Dependent Policies

We decentralize constrained efficient equilibria with size-dependent taxes on bank gross returns. In our procedure, we want the decentralized solution to converge to the social planner's allocations both in terms of aggregates and in the banking cross-section. We impose on the decentralized solution optimal endogenous aggregates states from the constrained efficient equilibrium above. Then, we iterate on a tax schedule until all policy functions yield idiosyncratic endogenous states (the deposit rate schedule) that are exactly consistent with the social planner's allocations. Specifically, we start with a guessed $\tau(j)$ for each grid point. We solve the problem of the financial intermediaries subject to the tax schedule and compute a new guess for $\bar{R}(j)$, and so on until convergence. We do not update the household's solution or run simulations in the intermediate step, because the aggregate endogenous states are fixed. The exact algorithm is described below

Algorithm - Regulated Market Equilibrium

1. Start with the solution to the decentralized equilibrium. Impose new aggregate endogenous states that are now permanently equal to the social planner's values: $\{K_{sp}, P_{sp}, \Lambda_{sp}\}$. Compute R^k once and do not change during the iteration.

2. Guess some initial values for the gross return tax $\tau(j)$
3. Solve the financial intermediation problem
   (a) Treat return taxes $\tau(j)$ as given
   (b) Use value function iteration. On each grid point, assume the leverage constraint binds.
   (c) Construct the Lagrange multiplier. If constraint indeed binds, proceed. If constraint is slack, solve the problem again using a numerical minimization routine
4. Calculate the new probability of bank default on each grid point using the newly computed policy functions. This gives new $\{\upsilon'(j), \bar{R}'(j)\}$.
5. Compare the policy function for net worth n' obtained in this iteration with the social planner's solution n'_{sp}. Use a bisection method for constructing a new candidate $\tau'(j)$. if $n'(j) > n'_{sp}(j)$ for any j, increase the tax rate on that grid point. Alternatively, the regulated solution is too small and we decrease the tax rate on that grid point.
6. Compare the new $\tau'(j)$ with the old $\tau(j)$. If the maximal squared error across all grid points is within the tolerance level, complete the program. Alternatively, update $\tau(j)$ and revert back to Step 3.