# A Macroeconomic Model with Heterogeneous Banks

Rustam Jamilov[†]
London Business School

This version: April 8, 2021
First version: March 28, 2020

**Abstract**

I study positively and normatively the role of bank heterogeneity in the macroeconomy. I build an empirically-motivated macroeconomic model with a banking sector that features uninsurable idiosyncratic rate of return shocks, endogenous markups, costly default, and endogenous entry. The framework highlights a *trilemma* for bank regulation: the government cannot simultaneously improve financial competition, efficiency, and stability. Three validated channels impact the transmission of policy regimes on the macroeconomy: an *economies of scale* channel from larger banks being more efficient, an *endogenous competition* channel from larger banks charging higher markups, and a *financial stability* channel from smaller banks facing shorter distance to default. The trilemma extends to deposit insurance schemes, heterogeneous capital requirements, the "too-big-to-fail" hazard, and optimal constrained efficient allocations. I discuss implications of the framework for the ongoing rise of banking concentration, emergence of fintech credit, targeted stabilization policies like bank-level bailouts and liquidity facilities, and intermediary asset pricing.

*Keywords*: Financial intermediaries, heterogeneous agents, market power, bank regulation, financial frictions, dynamic equilibrium models. *JEL Codes:* E44, G20, G21, G28, G32, G38

[†]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 thank Michele Andreolli, Fernando Broner, Gabriel Chodorow-Reich, Nuno Coimbra, Dean Corbae, Briana Chang, Itamar Drechsler (discussant), Ester Faia, Miguel Faria-e-Castro, Andrea Ferrero, Luca Fornaro, Xavier Gabaix, Sigurd Galaasen, Francois Gourio, Francisco Gomes, Veronica Guerrieri, Kyle Dempsey, Tarek Hassan, Zhiguo He, Kilian Huber, Luigi Iovino, Ragnar Juelsrud, Diego Kaenzig, Anil Kashyap, Nobu Kiyotaki, Ralph Koijen, Joseba Martinez, Michael MacMahon, Frederic Mishkin, Tommaso Monacelli, Simon Mongey, Plamen Nenov, Tsvetelina Nenova, Elias Papaioannou, Pascal Paul, Franck Portier, Albert Queralto, Morten Ravn, Ricardo Reis, Kenneth Rogoff, Petr Sadlacek, Andrew Scott, Vania Stavrakeva, Vincent Sterk, Kjetil Storesletten, Ludwig Straub, Paolo Surico, Jenny Tang, John Vickers, Annette Vissing-Jorgensen, Randall Wright, Francesco Zanetti, Tong Zhang and participants at various conferences and seminars for helpful feedback. I also thank Andrea Pasqualini for sharing his data on bank markups. I am grateful to the UniCredit Foundation, 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

Is there a trade-off between competition, efficiency, and stability in the modern banking system? This question has remained at the core of policy-relevant empirical and theoretical research on banking over the past decades (Corbae and Levine, 2018). In this paper, I argue that we should think of these three dimensions through the lenses of a "trilemma": any policy intervention that enhances one of these structural facets necessarily exacerbates one or more of the remaining two. This is a simple and novel generalization of the canonical financial competition-stability debate in a world where banks also differ systematically in their market power. The trilemma offers a fresh perspective on classic issues in bank regulation like the "too-big-to-fail" hazard, deposit guarantee schemes, and capital requirements. It also has immediate implications for new trends in banking such as the rise of concentration and emergence of fintech-intermediated credit. Furthermore, the trilemma can guide practical implementation of unconventional monetary and fiscal tools like targeted bailouts and liquidity facilities.

The banking industry trilemma arises naturally in a tractable macroeconomic model with a financial intermediation sector that is consistent with the following four motivating facts:

**Fact 1**: *The banking industry is highly concentrated.* Moreover, the industry is becoming more concentrated over time. This is true for the U.S. as well as for the Euro area (Corbae and D'Erasmo, 2020b; Constancio, 2016). As of the end of 2020, the 10 largest banks in the U.S. controlled roughly 60% of the nationwide loan market. We need a quantifiable framework that can generate reasonable cross-sectional dispersion and concentration of bank size.

**Fact 2**: *There are economies of scale in lending; larger banks are more efficient than smaller banks.* Multiple empirical studies have confirmed presence of either cost- or productivity-driven economies of scale in banking (Wheelock and Wilson, 2012, 2018; Berger and Mester, 1997; Berger and Hannan, 1998). As a bank's balance sheet grows, both marginal and fixed costs begin to shrink relative to assets under management. Economies of scale is also the cornerstone of core principles in canonical banking theory such as delegated monitoring (Diamond, 1984). A realistic quantitative model must therefore be able to generate heterogeneity in intermediary productive or lending efficiency.

**Fact 3**: *Bank markups are heterogeneous; larger banks charge higher loan markups than smaller banks.* This relatively novel stylized fact has appeared in the works of Corbae and D'Erasmo (2020a) and Pasqualini (2021). Authors apply a variant of the production-function approach that De Loecker et al. (2020) have popularized for the study of market power and find that bank markups are concentrated in the right tail of the size distribution. Elsewhere, Benetton (2021) in a structural model of the UK mortgage market also finds that larger banks charger higher loan markups. We thus need a model with imperfect financial competition and variable markups.

**Fact 4**: *Bank default risk declines with bank size; default is costly for the real economy.* Using quarterly bank-level data on U.S. commercial banks we will establish that exit risk is heavily concentrated in the left tail of the bank size distribution. The literature on the social costs of financial crises is vast and some of the notable contributions include Reinhart and Rogoff (2009), Schularick and Taylor (2012), Romer and Romer (2017), and Laeven and Valencia (2018). Literature consensus seems to be that financial crises, especially systemic banking crisis episodes, lead to considerable, prolonged declines in economic activity, financial intermediation, and consumer welfare. We need a model where heterogeneous banks face endogenous exit risk that is costly for society.

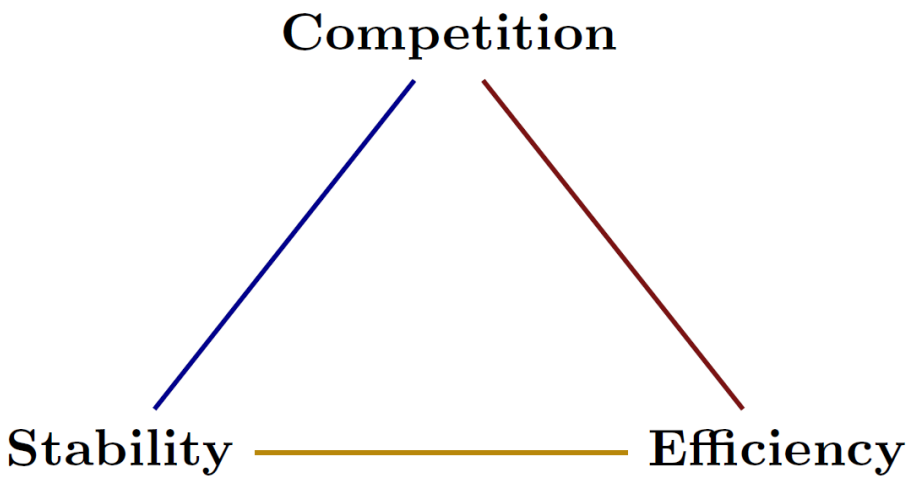
To formalize these facts into a coherent framework, I build a parsimonious dynamic general equilibrium macroeconomic model with heterogeneous banks. There are four main buildings blocks to this quantitative theory. First, we start with a stripped-down version of the workhorse representative-bank macroeconomic environment of Gertler and Kiyotaki (2010) and Gertler and Karadi (2011) and nest them as a special case. Second, banks engage in monopolistic competition in the credit market with non-CES demand, as in Kimball (1995). This is a tractable way to engineer variable loan markups that increase with relative balance sheet size. This setup nests the Dixit and Stiglitz (1977) (CES) aggregator as a special case and has been applied widely in the literature on monopolistic competition with non-financial firms (Klenow and Willis, 2016; Midrigan et al., 2018; Baqaee et al., 2021).[1] Third, banks face partially uninsurable idiosyncratic rate of return risk in the spirit of Benhabib et al. (2019). This assumption is motivated, among others, 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. Uninsurable idiosyncratic shocks create an exogenous bank net worth fluctuation problem analogous to the canonical Bewley-Huggett-Aiyagari-Imrohoglu environment (Bewley, 1977; Huggett, 1990; Aiyagari, 1994; Imrohoglu, 1996).[2] Finally, idiosyncratic risk is a source of insolvency risk for banks, who can default on their short-term debt obligations. Default is costly and the cost increases with the size of the bank.

The calibrated model is validated by reproducing all four facts that we document above. First, the model generates realistic, concentrated stationary distributions of bank net worth, assets, book and market leverage ratios, markups, relative prices, default probabilities, and deposit rates. This reconciles Fact 1. Second, in equilibrium smaller banks (a) are "unlucky" with a bad history of idiosyncratic shocks, (b) have shorter distance to default, and (c) face higher equilibrium interest rates on short-term debt. All three factors contribute to smaller banks facing higher marginal costs. Stationary marginal cost heterogeneity determines the *economies of scale channel* - bigger banks

[1] To the best of my knowledge, mine is the first attempt to apply this highly effective modelling technique to the case of monopolistically competitive lending markets.

[2] Embedding the "Bewley problem" into the banking sector of a dynamic general equilibrium framework is a second contribution of the paper.

Figure 1: **The Banking Industry Trilemma**




Notes: Figure visualizes the competition-efficiency-stability trade-off that arises in the model. Competition stands for the average equilibrium bank markup. Efficiency is the average marginal cost or the inverse of the credit supply elasticity of banks. Stability is the average probability of bank default due to insolvency.

have a higher lending capacity because they are endogenously more cost- and productive-efficient. This result is consistent with Fact 2. Third, because of the Kimball aggregator, high-net-worth banks choose to charge higher markups. This is the endogenous *competition channel* that replicates Fact 3. Fourth and finally, larger banks face a lower equilibrium probability of default but their endogenous exit carries a greater cost for the economy. This is the *financial stability channel*, which is in line with Fact 4.

We can now discuss the key unifying theme of all quantitative results - the *banking industry trilemma*. Figure 1 helps to visualize this result. In the stationary equilibrium, we obtain simultaneously that bigger banks are more efficient, default less often, but charge higher markups than smaller banks. There is therefore a trade-off between the economies of scale, endogenous competition, and financial stability channels. Any regulation-induced reallocation of credit towards the right tail of the bank size distribution improves aggregate stability and efficiency but reduces competition via rising markups. On the other hand, any reallocation of credit towards low net-worth banks decreases the average markup but worsens efficiency and stability. All in all, no singular regulatory intervention can improve all three dimensions of the banking system simultaneously as long as the efficiency-competition-stability trade-off is part of the economic environment. It is a policy trilemma.

I illustrate the workings of the trilemma on several classic topics in bank regulation. First, regulatory capital requirements that increase with bank leverage can improve financial stability by reducing the aggregate leverage ratio and systemic risk. However, this intervention meddles with the banks precautionary lending motive and their ability (and desire) to grow. As a result, the

regulated economy is less efficient as aggregate lending falls and costs rise. Second, introducing full deposit insurance generally has a positive effect on lending and growth but a large negative effect on stability and a positive effect on markups. Third, an extension of the model with the "too-big-to-fail" subsidy causes all macroeconomic aggregates like lending and production to increase while systemic fragility goes up.[3] Finally, constrained-efficient allocations and optimal heterogeneous bank taxation, which fully internalize the impact of all bank choices (assets, debt, markups) on aggregate prices and returns, improve gross welfare but severely worsen systemic financial stability. When default is sufficiently costly ex-post, net welfare effects could be negative.

In the rest of the paper, I apply the framework and the trilemma to several old and new issues in macro-banking. First, the global banking industry is becoming more and more concentrated. My theory predicts that this permanent "granularity" shock will make the banking system more efficient, stable, but less competitive. Second, the worldwide share of fintech and bigtech in financial intermediation is growing rapidly (Claessens et al., 2018). My framework predicts that the emergence and rise of fintech credit will lead to economic growth, a significant increase in the number of active banks, but ultimately a decline in financial stability since the economy would be populated with too many small and risky intermediaries which lack the scale to withstand idiosyncratic uncertainty.

The model has two additional auxiliary implications that could be useful in their own right. First, on the implementation of various unconventional, bank-level stabilization policies that have become very popular since the 2007-2008 Great Financial Crisis. I find that policies that shift relative prices or marginal costs, such as targeted liquidity facilities, are more effective when applied only to small banks. On the other hand, aggregate efficiency gains from unanticipated targeted equity injections ("bailouts") increase with the size of the impacted intermediary. Second, the model offers a fresh perspective on intermediary asset pricing with heterogeneity. My framework can generate a sizable unconditional risk premium due to heterogeneity in liquidity and default risk premia. Moreover, the distribution of bank size is procyclical, implying that liquidity and default risk premia are countercyclical, thus generating endogenous counter-cyclicality in the aggregate risk premium.

**Literature.** This paper contributes to several literature strands.

First, I build on a long literature that studies the tradeoffs between financial competition, stability, and growth (efficiency). One view is that competition reduces franchise values of the banks and induces more risky behavior. (Keeley, 1990; Hellman et al., 2000; Repullo, 2004; Beck et al., 2006). There is also an alternative view that riskiness of loans correlates with the level of the interest rate. As a result, greater competition may reduce default risk (Boyd and Nicolo, 2005;

[3]This finding is consistent with the idea that the TBTF subsidy makes private leverage choices of individual banks strategic complements (Farhi and Tirole, 2017).

Martinez-Miera and Repullo, 2010). My contribution is to reconsider these classic trade-offs in a novel general equilibrium environment where bank markups are endogenously heterogeneous and scale-dependent.[4]

Second, several studies have emphasized the role of financial heterogeneity and rely, like my model does, on some form of idiosyncratic risk and ex-post heterogeneity. Such papers include Corbae and D'Erasmo (2020a), Bianchi and Bigio (2020), Rios Rull et al. (2020). Rios Rull et al. (2020) study countercyclical capital buffers in a partial-equilibrium setting with idiosyncratic bank default risk. Bianchi and Bigio (2020) study competitive banks' liquidity management problem in a model with idiosyncratic deposit withdrawal shocks. Corbae and D'Erasmo (2020a) study oligopolistically competitive banks that are subject to idiosyncratic shocks on the liability side of the balance sheet. My main contributions are twofold. First, I study market power and heterogeneity stemming from the asset side of the banks balance sheet with a highly flexible monopolistic financial competition setup that can accommodate both constant (CES) and variable (Kimball) markups. Second, I explore normative implications in a realistic but complex environment with incomplete markets, variable bank markups, and default risk.

Third, my model is related to the literature that introduces *ex-ante* heterogeneity among financial intermediaries. Coimbra and Rey (2019) develop a general equilibrium model with ex-ante heterogeneity in intermediary value-at-risk constraints and endogenous financial stability. Their model features, like ours, dynamic intensive and extensive margins of bank risk-taking. My 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 and balance sheet characteristics. Second, my model departs from the assumption of perfect competition in bank lending. This channel delivers rich ex-post variation in markups and relative prices.[5]

Finally, this paper contributes to a long-running literature that introduces credit frictions and financial intermediaries into general equilibrium macroeconomic models. I build on the workhorse macro-banking setup of Gertler and Kiyotaki (2010) and Gertler and Karadi (2011), whom my model nests as special cases. A very incomplete list of salient equilibrium models with a financial sector includes Gromb and Vayanos (2002), Brunnermeier and Pedersen (2009), Adrian and Shin (2010, 2014), Jermann and Quadrini (2013), Brunnermeier and Sannikov (2014), He and Krishnamurthy (2013), Nuno and Thomas (2017), Gertler et al. (2016, 2020), Fernandez-Villaverde et al. (2019), Lee et al. (2020), Bigio and Sannikov (2021) etc. These frameworks largely abstract

[4]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). 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.

[5]Other papers that develop models with elements of financial heterogeneity include Boissay et al. (2016), Begenau and Landvoigt (2020), Stavrakeva (2020), Begenau et al. (2020), Goldstein et al. (2020), Dempsey (2020).

from distributional considerations in the financial sector and work with a representative financial intermediary/entrepreneur/arbitrageur whose relative scale generally cannot be pinned down.

**Outline.** The rest of the paper is structured as follows. Section 2 provides three motivating facts on the cross section of U.S. banks. Section 3 lays out the model. Section 4 discusses how we take the model to the data. Section 5 presents the main quantitative analysis of the paper. Section 6 concludes. Finally, the Online Appendix contains additional applications, derivations, extensions, numerical algorithms, and data description.

# 2 Cross-Sectional Banking Data

In this section I document three motivating cross-sectional facts on bank balance sheet size, leverage, markups and exit risk. We will use these facts in order to keep the model accountable.

## 2.1 Leverage

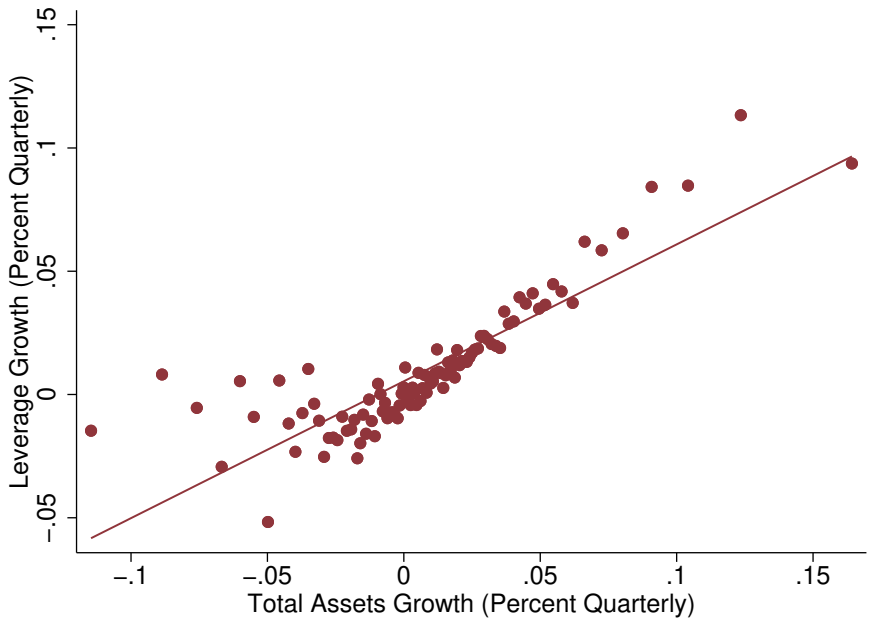
What is the relationship between bank leverage and balance sheet size? We answer this question using the data from Consolidated Reports of Condition and Income (the Call Reports). Every insured bank in the U.S. must submit these reports to the Federal Reserve on a quarterly basis. Following Adrian and Shin (2010) we measure balance sheet size with total assets. Leverage is defined as the ratio of total assets to total equity. I focus on the 2010:q1-2019:q4 period, based on which the model in the next section will be calibrated. Table 5 in Section G of the Online Appendix describes how every variable is defined and constructed.

Figure 2 plots a binned scatter plot between total assets and leverage. Both are quarterly growth percentages. The x-axis is binned into 100 percentiles of the assets growth distribution. For each bin, we compute the average of quarterly leverage growth within that bin. We see a clear positive correlation between assets and leverage, implying "procyclical" leverage in the words of Adrian and Shin (2010). The positive relationship between leverage and size of financial firms has been recently highlighted in Coimbra and Rey (2019). Gopinath et al. (2017) document a similar fact on a large sample of non-financial firms.

## 2.2 Lending Markups

What is the relationship between bank balance sheet size and lending markups? Estimation of market power in the banking industry is a relatively new literature. The main empirical approach involves the production function estimation methodology popularized by De Loecker et al. (2020). Variants of this approach have been recently applied to the case of U.S. banks by Corbae and

Figure 2: **Bank Assets and Leverage**




Notes: Binned scatter plot of bank assets and leverage growth in the data. The x-axis is binned into 100 percentiles of the distribution of quarterly assets growth. The y-axis is the bin-specific average of quarterly leverage growth. Leverage is defined as book assets over book equity. The leverage and assets growth distributions have been pre-trimmed at the 1% and 99% levels. Quarterly data is pooled over 2010:q1-2019q4. Source: Call Reports.

D'Erasmo (2020a) and Pasqualini (2021) who also estimates markdowns on the liability side. Elsewhere, Wang et al. (2020) and Benetton (2021) obtain bank markups using structural estimation.

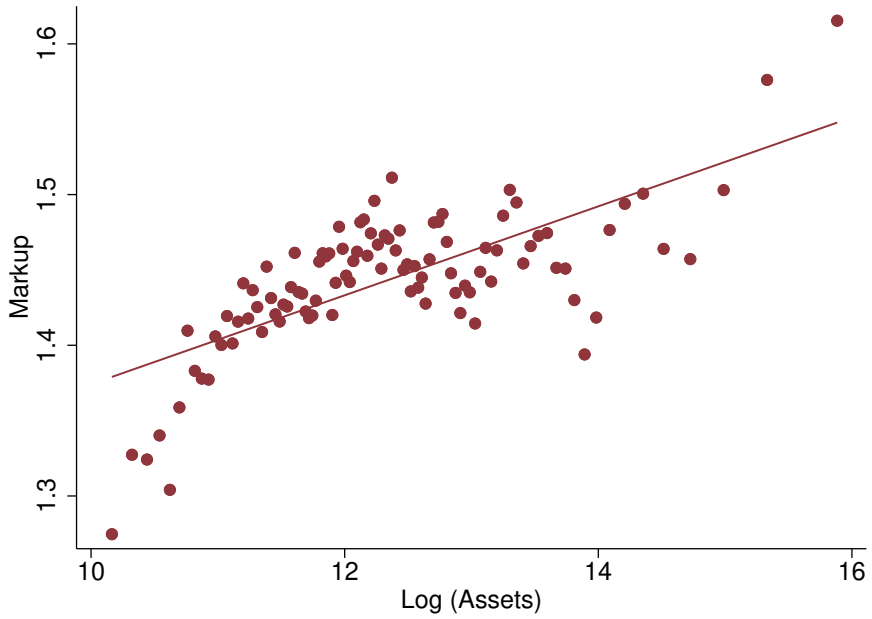
In the context of the present paper, we borrow estimates of markups from Pasqualini (2021), to whom I refer the interested reader for details related to methodology. Figure 3 plots a binned scatter plot between bank assets and absolute markups. Annual data is pooled over 2010-2019 and is based on the Call Reports. The x-axis is binned into 100 percentiles of the assets distribution. For each bin, we compute the bin-specific unweighted average of markups. From the plot we see a clear, strong positive correlation between assets and markups. This qualitative relationship is emphasized also by Corbae and D'Erasmo (2020a) and Benetton (2021). It appears that larger banks charge higher markups over their marginal costs.

## 2.3 Exit Risk

What is the relationship between bank balance sheet size and exit risk? I impute bank exit risk using two separate approaches.[6] The first approach relies on bank balance sheet and income

[6]As we will see in the model, I won't be differentiating between outright exit from being acquired by another institution. The secondary mergers and acquisitions market will not be explicitly modelled. As a result, we do not

Figure 3: **Bank Assets and Markups**




Notes: Binned scatter plot of bank assets and lending markups in the data. The x-axis is binned into 100 percentiles of the distribution of Log (Assets). The y-axis is the bin-specific average of lending markups. Markups and assets distributions have been pre-trimmed at the 1% and 99% levels. Annual data is pooled over 2010-2019. Markups are from Pasqualini (2021) and assets are from Call Reports.

statement data from the Call Reports. We begin by constructing an indicator variable $\text{Exit}_{it}$ for each bank $i$ and quarter $t$ which takes the value of unity if the bank exits in quarter $t+1$. We then run the following logit regression of $\text{Exit}_{it}$ on assets.

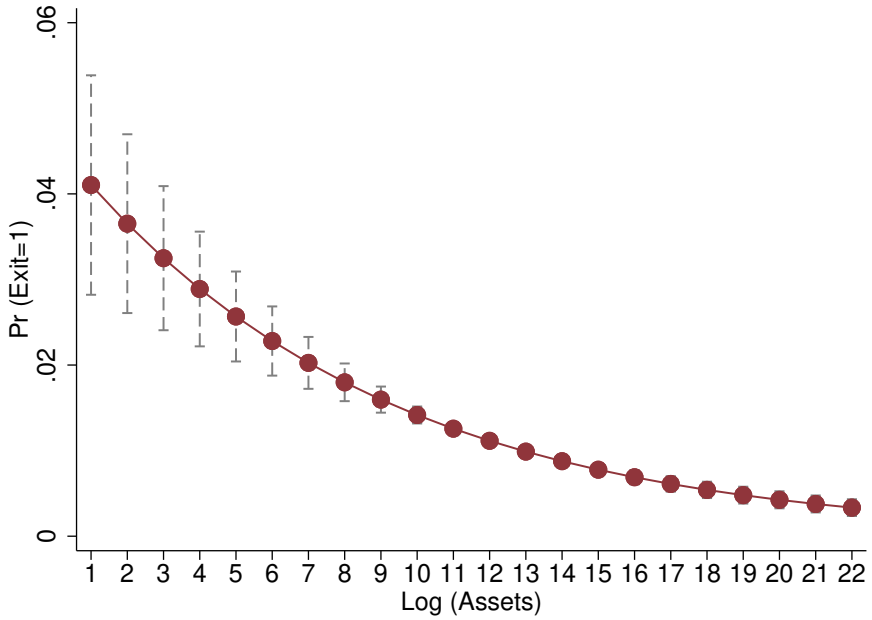
$$\Pr\{\text{Exit}_{it} = 1\} = \alpha_0 + \alpha_1 \text{Log (Assets)}_{it} + \mu_t + \epsilon_{it} \tag{1}$$

Where $\mu_t$ is a quarterly fixed effect. It captures the idea that the likelihood of exit is potentially aggregate state-dependent. Standard errors are robust to heteroskedasticity and serial correlation. When Log (Assets) are held to their mean value, the predicted probability of exit is 1.074%. We can also compute the probability of exit conditional on different asset values. Figure 4 plots the margins plot from our logit regression and depicts all the point and interval estimates of the conditional probability. The predicted probability of exit ranges from roughly 4% for the first percentile of the asset distribution to 0.33% for banks in the top percentile. This result is quite intuitive: smaller banks are more exposed to exit risk.

As a robustness check, I also impute the likelihood of bank exit based on the Laeven and

differentiate between these two sources of exit in our data treatment.

Figure 4: **Bank Assets and Exit Risk**




Notes: Marginal point and interval estimates from the logit regression of the indicator for bank exit $\text{Exit}_{it}$ on Log (Assets) with a time fixed effect and standard errors that are robust to serial correlation and heteroskedasticity. The margins plot shows conditional probabilities of exit at various values of Log (Assets). Quarterly data is pooled over 2010:q1-2019q4. Source: Call Reports.

Valencia (2018) database of banking crises from around the Globe.[7] According to the authors, there have been 107 unique banking crises events over the past 48 years across 165 countries.[8] That makes the unconditional probability of a crisis equal to roughly 1.35%, which is in the ballpark with the 1.074% that we computed from the Call Reports.

Because the model will feature *costly* bank default, we also require proxies for the real costs of banking crises. For this purpose, we again rely on the empirical findings in Laeven and Valencia (2018). Authors estimate that output losses around systemic banking crises historically average around 7.6% of the difference between potential and actual GDP per year. They also find that these losses tend to be larger in advanced economies (11.67%), which are more financially sophisticated, than in emerging countries (4.67%). I will use this evidence to motivate that default of larger intermediaries in the model is more costly ex-post than of smaller banks.

[7] In the model, bank exit will be synonymous with financial crises.

[8] I focus exclusively on incidents of banking crises only, excluding concurring sovereign or exchange rate crises.

# 3 A Model with Heterogeneous Intermediaries

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

## 3.1 Overview

Time is discrete and infinite. The economy consists of a representative household, a continuum of financial intermediaries that are ex-ante identical but ex-post heterogeneous, a representative final goods producer, and a representative capital goods producer. The household is risk-averse, supplies labor inelastically to the final goods firm in exchange for a competitive wage, and saves intertemporally through one-period bank deposits. The deposit market is perfectly competitive and there is no deposit insurance in the baseline economy; we will introduce it in Section 5.2.

Banks accumulate own net worth, acquire deposits from households to whom they pay the equilibrium deposit rate the following period, cover non-interest expenses that are non-linear in assets under management, and perform two investment activities. First, banks invest into claims on zero-profit capital goods firms who produce aggregate capital.[9] Competition is monopolistic and demand is non-CES (Kimball).[10] After the capital stock is produced and priced, banks immediately lend it competitively to the final goods producer in return for realized returns on capital the following period.[11] In addition to the systematic return, each bank receives a bank-specific idiosyncratic return draw. These shocks are persistent and not perfectly insurable. Along the extensive margin, banks exit due to exogenous and endogenous reasons. Upon exogenous exit, banks pass on all remaining net worth to households, which motivates our constant dividend payout rule. Endogenous exit is due to default (fundamental insolvency). Finally, entry is exogenous in the baseline economy. Entry and the mass of active banks are endogenized in Section A.2 of the Online Appendix where we discuss fintech-intermediated credit.[12]

[9] We abstract from bond financing which is another major source of external funding for firms (De Fiore and Uhlig, 2011).

[10] Why don't we set up an oligopolistic credit market? There are at least three reasons. First, although the number of active banking franchises in the U.S. and Europe is dwindling, there are still more than 5,000 active commercial banks in the U.S. at the time of writing. A monopolistic competition structure feels much more appropriate given the number of agents in the market. Second, it is more reasonable that individual banks do *not* internalize the impact of their private choices on aggregate outcomes. Banks in our model are "atomistic" because of the monopolistic competition assumption. But they are still "granular" in the sense that concentration of the distribution matters for macroeconomic outcomes (Gabaix, 2011). Third and finally, tractability. There is little more that an oligopoly model can give us that the monopolistic competition block with variable markups cannot. A flexible two-parameter departure from perfect competition cannot be taken for granted in an environment with incomplete markets and uninsurable idiosyncratic shocks like ours.

[11] Ownership of the capital stock is in pro rata terms.

[12] In Section E.3 of the Online Appendix we show how the baseline economy can be extended to feature two sectors that are heterogeneous in the degree of financial competition.

An important advantage of the model is that any of the two essential building blocks - monopolistic financial competition and uninsurable idiosyncratic bank return shocks - can be shut down without affecting the other. In other words, we can analyze an economy with heterogeneity but perfect competition, monopolistic competition but homogeneity, or any re-calibrated combination in between.

## 3.2 Technology

**Final Good Production** The final good is produced from aggregate capital and labor using a Cobb-Douglas technology:

$$Y_t = AK_t^{\alpha} L_t^{1-\alpha} \tag{2}$$

where $0 < \alpha < 1$. Wages ($W_t$) and returns to capital ($R_t^k$) are competitive and follow directly from the firm's optimization problem:

$$R_{t+1}^k = \frac{A\alpha K_{t+1}^{\alpha-1}}{P_t} \qquad W_t = (1-\alpha)AK_t^{\alpha} \tag{3}$$

where A is aggregate productivity and $P_t$ is the price of capital. Capital depreciates fully every period after it's used in production.

**Capital Good Production** A representative, perfectly competitive capital producing firm begins the period with no equity and issues claims to banks in return for the aggregate capital bundle. The firm makes zero profits.

The capital good $K_t$ is produced from a bundle of financial varieties $k_t(j)$ for $j \in [0, 1]$. Financial varieties are intermediated by the banking sector and assembled with the *Kimball* aggregator:

$$K_t = \int_0^1 \Upsilon\left(\frac{k_t(j)}{K_t}\right) dj \tag{4}$$

where the function $\Upsilon(x)$ is increasing, concave, and satisfies $\Upsilon(1) = 1$. It can be shown that the Dixit-Stiglitz aggregator is a special case with $\Upsilon(x) = x^{\frac{\theta-1}{\theta}}$, where $\theta > 1$ is the constant elasticity of substitution.

The maximization problem of the capital goods firm is:

$$\max_{k_t(j)} \left[P_t K_t - \int_0^1 p_t(j)k_t(j)dj\right]$$

subject to technology 4. This yields a demand function for bank funds:

$$p_t(j) = \Upsilon'\left(\frac{k_t(j)}{K_t}\right) Z_t \tag{5}$$

where

$$Z_t := \left(\int_0^1 \Upsilon'\left(\frac{k_t(j)}{K_t}\right)\frac{k_t(j)}{K_t}dj\right)^{-1} \tag{6}$$

is the *Kimball demand index*. In the Dixit-Stiglitz special case, $Z_t = \frac{\theta}{\theta-1}$, and (5) reduces to $p_t(j) = \left(\frac{k_t(j)}{K_t}\right)^{\frac{-1}{\theta}} P_t$.

**Discrete Choice Microfoundation** It is possible to theoretically underpin the monopolistic credit demand system above using discrete choice theory where each borrower chooses both the size of the loan and the bank/variety to borrow from (McFadden, 1984). The approach generalizes the case of a representative capital goods producer to a large number of borrowers that are heterogeneous in their preferences for individual banks. In other words, there are firm-bank fixed-effect shocks. These shocks are cross-sectionally correlated and the degree of correlation maps into the constant elasticity of substitution $\theta$. Section E.1 of the Online Appendix provides a detailed guide for the analytically convenient case of $\epsilon = 0$.

Market power at the level of a bank can now be viewed as being isomorphic to consumer (firms, in this case) preferences for financial services that are not perfect substitutes across banks. Even if a particular bank chargers higher prices, it can still remain in business if borrower-bank-specific preference shocks are sufficiently diverse. 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 working with this representative-firm approximation of the more sophisticated heterogeneous-firms environment that is understood to be operating in the background.

## 3.3 Banks

The general credit demand system in (4)-(6) 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 assets k(j), deposits d(j), and price of claims p(j) while respecting the balance sheet constraint:

$$d_t(j) + n_t(j) = p_t(j)k_t(j) \tag{7}$$

Every bank faces non-interest expenses $\frac{1}{\zeta_1}k_t(j)^{\zeta_2}$ where parameter $\zeta_2$ can help govern the degree of non-linearity and scale-variance. Section D.1 in the Online Appendix demonstrates how whenever $\zeta_2 \neq 1$ aggregate state-dependency on n(j) is achieved, i.e. bank characteristics matter for aggregation.

When choosing the size of the balance sheet, banks can borrow deposits d(j) from the household, subject to the bank-specific interest rate $\bar{R}_t(j)$ that will be determined in general equilibrium. 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 $R_t^k$, which is common to all j, and an idiosyncratic component $\xi_t(j)$ which is specific to each j:

$$R_t^T(j) = \kappa\xi_t(j) + (1-\kappa)R_t^k \tag{8}$$

Where $0 < \kappa < 1$ is a parameter that governs the ability to hedge idiosyncratic shocks. We discuss a possible microfoundation for the $R^T(j)$ formulation in Section E.2 of the Online Appendix. The idiosyncratic return, $\xi \in \Xi$, follows an AR(1) process:

$$\xi_t(j) = (1-\rho_\xi)\mu_\xi + \rho_\xi\xi_{t-1}(j) + \sigma_\xi\epsilon_t(j) \tag{9}$$

Analogously, let the finite state Markov representation of (9) be $\mathbf{G}(\xi_{t+1}, \xi_t)$. We can now state the law of motion of bank-level net worth:

$$n_{t+1}(j) = R_{t+1}^T(j)p_t(j)k_t(j) - \bar{R}_t(j)d_t(j) - \frac{1}{\zeta_1}k_t(j)^{\zeta_2} \tag{10}$$

Following Gertler and Karadi (2011) and Gertler and Kiyotaki (2010), there is a moral hazard problem in the banking sector. Banks have an incentive to divert franchise assets with the ability to divert no more than a fraction $\lambda$ 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 divert exactly equals the value of the franchise. This yields the following incentive constraint that puts a limit on bank leverage.

$$\lambda p_t(j)k_t(j) \leq V_t(j) \tag{11}$$

where $V_t(j)$ is the franchise value of the intermediary, to be defined below. Each bank in the economy can default with an endogenous probability $\upsilon(j)$, which is taken as given and determined in equilibrium. Default risk is due to fundamental insolvency, i.e. when net worth at normal market prices is non-positive.:

$$\upsilon_t(j) = \Pr\Big(n_{t+1}(j) \leq 0\Big) \tag{12}$$

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 will be concentrated in the *left* tail of the stationary bank net worth distribution, which is in line with our empirical analysis in Section 2.

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

$$\eta_{t+1}(n_{t+1}, \xi_{t+1}) = \Phi(\eta_t) \tag{13}$$

We define $\Phi$ 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 $\sigma$ the incumbent may exit involuntarily, in which case its net worth gets transferred lump sum to the household. The banker discounts the future by adopting and augmenting the household's stochastic discount factor $\Lambda$, which is determined in equilibrium and defined when we discuss the household problem. Each banker takes as given aggregate quantities $\{K, D, N\}$, prices $\{P, R^k\}$, the cross-sectional distribution $\eta$ and its law of motion $\Phi$, bank-specific deposit rates $\bar{R}$ and portfolio returns $R^T$. Each bank solves:

$$V(\mathbf{s}) = \max_{\{k,p,d\}} \left\{ \mathbb{E}_{\mathbf{s'|s}} \left[ \Lambda' \Big( (1-\sigma) n' + \sigma V(\mathbf{s}') \Big) \right] \right\} \tag{14}$$

$$\text{s.t. conditions 4-13.}$$

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:

$$\max_{\phi} \left[ \mu_a \phi + \nu_a \right] \tag{15}$$

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, $\nu_a = (1-\upsilon)\tilde{\Lambda}' \left[ \bar{R} - \frac{\frac{1}{\zeta_1} k_t(j)^{\zeta_2}}{n} \right]$ is the cost of liabilities. In both instances, $\tilde{\Lambda}' = \Lambda \left( 1 - \sigma + \sigma V(\mathbf{s}') \right)$ is the augmented household marginal rate of substitution.

Section D.2 of the Online Appendix shows that the solution to the above problem, while taking all aggregate quantities and equilibrium prices as given, yields the following relative price rule:

**Proposition 1** (Markups and Marginal Costs Decomposition)**.**

$$\frac{p(j)}{P} = \mu(x)\frac{k(j)^{\zeta_2-1}}{R^T(j) - \bar{R}(j)} \tag{16}$$

where $\mu(x)$ is a markup function, which potentially depends on relative size $x = \frac{k(j)}{K}$, and $\frac{k(j)^{\zeta_2-1}}{R^T(j)-\bar{R}(j)}$ the endogenous marginal cost. In the two paragraphs that follow, we zoom in on these two sources of bank heterogeneity in the model: markups and marginal costs.

## 3.4 Variable Markups

For the baseline case with endogenously variable bank markups, I use the Klenow and Willis (2016) specification for $\Upsilon(x)$:

$$\Upsilon(x) = 1 + (\theta - 1)\exp\frac{1}{\epsilon}\epsilon^{\frac{\theta}{\epsilon}-1}\left[\Gamma\left(\frac{\theta}{\epsilon}, \frac{1}{\epsilon}\right) - \Gamma\left(\frac{\theta}{\epsilon}, \frac{x^{\frac{\epsilon}{\theta}}}{\epsilon}\right)\right] \tag{17}$$

where $\epsilon \geq 0$ is a parameter that governs variation in the *superelasticity* $\frac{\epsilon}{\theta}$ and $\Gamma(s, q)$ is the upper-incomplete Gamma function:

$$\Gamma(s, q) := \int_q^\infty t^{s-1}\exp^{-t} dt \tag{18}$$

The CES aggregator is a special case of (17) when $\epsilon = 0$. With the Klenow-Willis specification, we have:

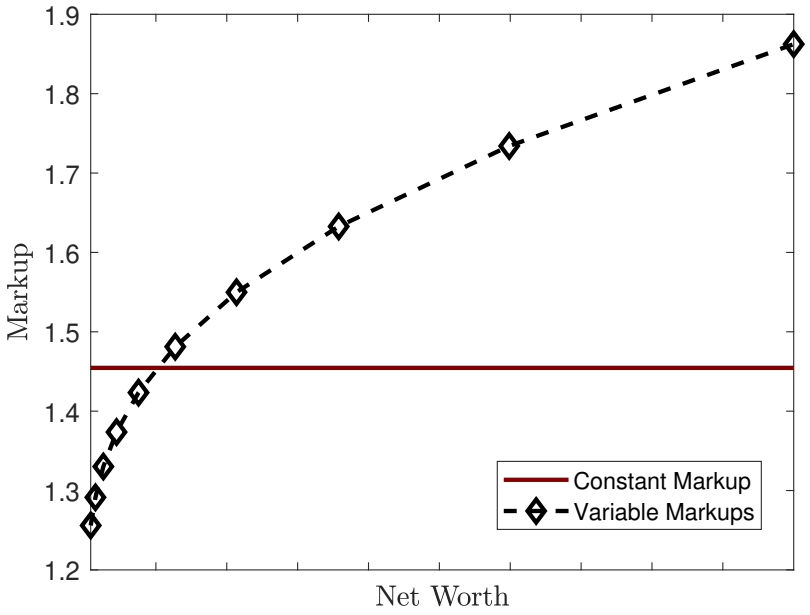
$$\Upsilon'(x) = \frac{\theta - 1}{\theta}\left(\exp\frac{1 - x^{\frac{\epsilon}{\theta}}}{\epsilon}\right) \tag{19}$$

The size-dependent elasticity is thus $\theta x^{-\frac{\epsilon}{\theta}}$. It can be seen clearly that the elasticity declines with relative size. This, in turn, implies the following markup function:

$$\mu(x) = \frac{\theta x^{-\frac{\epsilon}{\theta}}}{\theta x^{-\frac{\epsilon}{\theta}} - 1} \tag{20}$$

As long as $\epsilon > 0$, banks with a higher relative quantity of assets on their books ($x = \frac{k}{K}$) will face a lower elasticity of substitution. This, in turn, induces larger banks to choose higher $\mu(x)$. When $\epsilon = 0$, the credit markup is constant and equals the usual $\mu = \frac{\theta}{\theta-1}$. Calibration of the superelasticity

Figure 5: **Bank Markups**




Notes: Absolute bank markups with the Kimball ("Variable Markups"') and CES ("Constant Markup") aggregators.

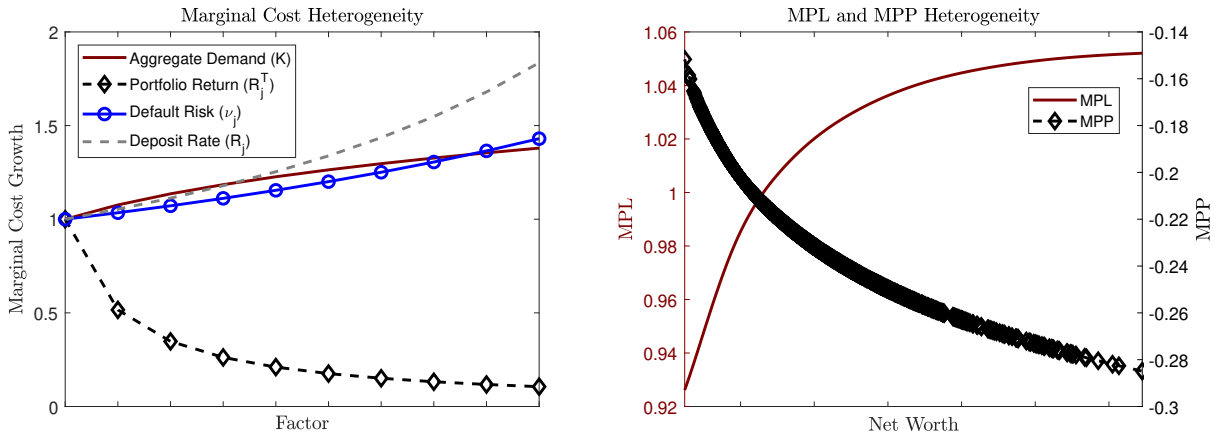
can be achieved in a simple way by varying $\frac{\epsilon}{\theta}$. When taking the model to the data, we will use the empirical cross-section of bank markups to deduce the two parameters.

Figure 5 illustrates the differences between Kimball-Klenow-Willis and Dixit-Stiglitz aggregators. Increasing $\epsilon$ makes the demand curve less "convex", everything else equal. Larger banks are in the area of relative satiation. Because they face lower substitution elasticities, they choose to charge higher markups since further reduction of relative prices does not induce the same desirable quantity effect. Also note how the shape of the markup function lines up exactly with the empirical relationship from Section 2. We will discuss the calibration strategy that achieves this in Section 4.

## 3.5 Marginal Costs and Economies of Scale

Cross-sectional bank heterogeneity in the model also runs through marginal costs. The marginal cost is a complex non-linear function of four key objects: total portfolio return $R^T(j)$, interest rate on deposits $\bar{R}(j)$, the scale effect in K, and the probability of default $\nu(j)$ (which affects marginal costs indirectly, through $\bar{R}(j)$). Note that dependency on aggregate demand K is only possible when non-interest expenses are not linear with respect to k(j), a condition we return to in the next section where we discuss scale variance. The effect of each of the four determinants on the marginal cost is summarized on the left panel of Figure 6. First, we observe that bank-level marginal

Figure 6: **Marginal Costs and Economies of Scale**




Notes: Left picture shows how bank marginal costs depend on aggregate demand, bank-level portfolio return draw, bank-level probability of default, and bank-level equilibrium deposit rate. Right picture plots marginal propensities to lend and marginal propensities to price as a function of bank net worth.

costs are increasing in aggregate demand. Greater demand 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. Finally, the marginal cost is decreasing in the portfolio return $R^T(j)$, which acts as a profitability shifter and is a source of all ex-post heterogeneity in balance sheet prices and quantities.

Marginal cost heterogeneity gives rise to economies of scale. In order to illustrate the mechanism in the cleanest possible way, we define two new objects: the marginal propensity to lend (MPL) and the marginal propensity to price (MPP). At the level of a bank, MPL(j) is constructed as the elasticity of assets k(j) to changes in net worth n(j). MPP(j) is 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):

$$\text{MPL} = \int_{\mathbf{B}} \frac{\partial \text{k(j)}}{\partial \text{n(j)}} \eta(\text{dn}, \text{d}\xi) \quad \text{MPP} = \int_{\mathbf{B}} \frac{\partial \text{p(j)}}{\partial \text{n(j)}} \eta(\text{dn}, \text{d}\xi) \tag{21}$$

The right panel of Figure 6 visualizes the MPL and MPP objects as functions of net worth. We see that MPP(j) and MPL(j) are inversely related, which is due to the Kimball demand function and negative correlation of assets and relative prices. In equilibrium, low-net-worth banks are those that (a) have a poor history of idiosyncratic return realizations, (b) have shorter distance to default, and (c) face higher equilibrium deposit rates. All three factors contribute to smaller banks facing higher marginal costs which, in turn, feeds into lower efficiency and lending capacity, which is summarized with a higher (lower) MPL (MPP). This is the economies of scale channel. For contrast, in the representative-bank counterfactual, the MPL and MPP distributions are flat and

correspond to the corresponding objects of the median intermediary. Depending on the extensive margin and the relative shares of very large and very small banks in the distribution, heterogeneity becomes important for the transmission of net worth shocks to aggregate lending and investment.

## 3.6 Entry and Exit

In the baseline version of the model, entry is exogenous. Upon entry, each bank receives a lump-sum endowment of net worth $n_0$ from the household and an idiosyncratic return draw that equals $\mu_\xi$. As mentioned before, we relax the assumption of exogenous entry in Section A.2 of the Online Appendix when we discuss the rise of fintech credit in an extension with endogenous entry. The incumbent intermediary is subject to two sources of exit risk: involuntary homogenous exit rate $\sigma$ and the endogenous probability of default $\nu(j)$, which is bank-specific. Default is due to fundamental insolvency, which occurs when n(j) is drawn down to 0.[13] Intermediary default is costly and results in ex-post efficiency losses that are measured in units of output. Default costs are potentially bank size-dependent. If a bank exits, the exiting bank's market will never be taken over by any of the incumbents.

## 3.7 Cross-Sectional Distribution of Banks

Denote $\bar{d}(n, \xi)$ a dummy variable which takes the value of unity if an individual bank exits in time t. Denote $M_t$ the mass of entering banks. This mass is predetermined and equals the mass of banks which exited due to the exogenous shock $(1-\sigma)$ or default. The mass of active intermediaries thus remains time-invariant. The distribution of banks in the economy evolves according to:

$$\eta'(n', \xi') = \sum_{\xi} G(\xi', \xi) \int \mathbb{1}_{\{(n,\xi)|K(n,\xi)\in \mathbf{B}\}} \times \mathbb{1}_{\{\bar{d}(n,\xi)=0\}} \eta(dn, d\xi) + M' n_0 \tag{22}$$

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

## 3.8 Households

The representative household is tasked with choosing the supply of deposits to each bank $b_t(j)$ and consumption $C_t$, subject to the standard balance sheet constraint:

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

[13] Bank runs are ruled out for tractability. For macroeconomic models with systemic bank runs see, for example, Uhlig (2010) or Gertler et al. (2020).

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

Where $\pi$ are any lump sum transfers or taxes. First order conditions for deposits yield the following equation:

$$\bar{R}_t(j) = \frac{1 - \upsilon_t(j)x_t(j)\mathbb{E}\left(R^T_{t+1}(j)\Lambda_{t+1}\right)}{\Big(1 - \upsilon_t(j)\Big)\mathbb{E}\left(\Lambda_{t+1}\right)} \tag{23}$$

Where $\Lambda_{t+1} = \beta\frac{u'(c_{t+1})}{u'(c_t)}$ is the stochastic discount factor and $u'(c) = c^{-\sigma_h}$. 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 $\phi$ the market leverage ratio, defined as before.

## 3.9 Stationary Industry Equilibrium

Credit market clearing requires:

$$K = \int_{\mathbf{B}} \Big(k(n, \xi)\Big)\eta(dn, d\xi) \tag{24}$$

Similarly, clearing the deposit market requires:

$$\int_0^1 b(j)dj = \int_{\mathbf{B}} \Big(d(n, \xi)\Big)\eta(dn, d\xi) \tag{25}$$

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 entering bankers:

$$Y = C + I$$

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 $V(\mathbf{s})$, optimal policies for bank capital investment $k(\mathbf{s})$ and deposit demand $d(\mathbf{s})$, household consumption C and deposit supply b(j), competitive wage W and capital $R^k$ pricing functions, the aggregate price of capital P, a marginal utility process $\Lambda$, and the menu of market-clearing deposit rates $\bar{R}(\mathbf{s})$ such that:

1. The household's choices $\{C, b(j)\}$ are optimal conditional on $\{W, \bar{R}(j)\}$
2. The banks choices $\{k, p, d, \mu\}$ are optimal conditional on $\{\Lambda, K, P, \bar{R}(\mathbf{s}), \eta\}$
3. Returns on factors of production are: $R^k = \frac{\alpha A K^{\alpha-1}}{P}$, $W = (1-\alpha)AK^{\alpha}$
4. Aggregate assets, deposits, and net worth $\{K, D, N\}$ are consistent with the cross-sectional distribution and the monopolistic credit demand system
5. The credit market clears as in (24). The deposit market clears as in (25)
6. The cross-sectional distribution evolves according to (22) and is consistent with bank-level demand functions

### 3.10 Numerical Algorithm

There are four basic steps in the computational algorithm. First, we must solve individual dynamic optimization problems of financial intermediaries (incumbent and new entrants, if entry is endogenous) and of the household. Because the banking sector is not scale invariant, individual bank characteristics $\{n(j), \xi(j)\}$ matter for aggregation. Second, banks face an occasionally binding constraint on leverage that could bind anywhere in the state space. Third, the market for deposit holdings must clear for each bank type. Finally, there are 3 key aggregate endogenous state variables that we need to pin down in general equilibrium: $\Lambda$, K and P. For K and P, we use a variant of the Maliar et al. (2010) stochastic simulation approach.

The algorithm is described in detail in Section F of the Online Appendix.

## 4 Taking the Model to the Data

In this section I discuss the parameterization strategy, moments that the model manages or fails to match, and some key cross-sectional patterns in the banking sector.

### 4.1 Parameterization

All chosen parameters are shown in Table 1. The model period is one quarter. We begin by describing standard macro parameters. We set the share of aggregate capital in production $\alpha$ to 0.36. The discount factor $\beta$ is chosen to target a steady-state annual risk-free rate of 1.60%. We assume log-utility in consumption.

For parameters in the banking block, we set the exogenous survival probability to $\sigma = 0.9$, which is consistent with a life expectancy of the average banker equalling 10 years, similarly to

Table 1: **Parameter Values**

| Parameter | Description | Value |
|---|---|---|
| | Macro | |
| $\alpha$ | Share of capital in production | 0.36 |
| $\beta$ | Discount factor | 0.996 |
| $\sigma_h$ | Risk aversion | 1 |
| | Banking | |
| $\sigma$ | Dividend payout ratio | 0.9 |
| $\omega$ | Share of divertible assets | 0.1 |
| $n_0$ | New banker endowment | 30% of N |
| $\frac{1}{\zeta_1}$ | Monitoring cost linear | 0.01 |
| $\zeta_2$ | Monitoring cost quadratic | 1.19 |
| | Monopolistic Credit Market | |
| $\theta$ | CES elasticity | 3.2 |
| $\frac{\epsilon}{\theta}$ | Superelasticity | 0.165 |
| | Idiosyncratic Bank Return Risk | |
| $\kappa$ | Fraction of portfolio exposed to idiosyncratic risk | 0.3 |
| $\rho_\xi$ | Serial correlation of idiosyncratic risk | 0.52 |
| $\sigma_\xi$ | SD of idiosyncratic risk | 0.085 |
| | Costly Bank Default | |
| $d_1$ | Default cost constant | 0.0511 |
| $d_2$ | Default cost linear | 0.0075 |

Gertler and Kiyotaki (2010). The fraction of divertible assets $\lambda = 0.1$ targets a steady state bank leverage ratio of roughly 7. Endowment of new entrants is set to 30% of average net worth $N$, which helps to achieve an empirically realistic average entry rate of 5% whenever entry is endogenous. Parameters that govern non-interest expenses $(\zeta_1, \zeta_2)$ are chosen to be consistent with empirical evidence on increasing returns to scale in banking while allowing the banking problem to remain concave (Wheelock and Wilson, 2018).[14]

Parameters of the monopolistic credit market block are chosen to hit two targets. First, based on the empirical evidence in Corbae and D'Erasmo (2020a) and Pasqualini (2021), the median markup of commercial banks in the U.S. over the 2010-2019 period was roughly 1.45, i.e. 145% over the marginal cost.[15] The average elasticity of $\theta = 3.2$ helps achieve a CES markup of 1.45.

[14]Our results do not rely on whether these costs are concave or convex, although convexity is much more computationally convenient. The knife-edge case of $\zeta_2 = 1$ is discussed in Section D.1.

[15]Using an approach that does not rely on production functions, Jamilov (2020) estimates branch-level loan price

Second, the superelasticity $\frac{\epsilon}{\theta} = 0.165$ generates the variable markup function seen on Figure 5.[16]

Parameters from the idiosyncratic return shocks block are chosen in order to match three facts. First, we motivate $\kappa$ as the portfolio share that banks allocate to the risky, shadow banking activities. Prior to the Great Financial Crisis the share of shadow banking activities in the broader financial intermediation sector of the U.S. was roughly 1/3 (Gorton and Metrick, 2010). Second, $\sigma_\xi$ is chosen in order to get the average probability of involuntary exit in line with the data. I described how those probabilities are estimated from the data in Section 2.3. Third, persistence $\rho_\xi$ is chosen in order to get the skewness of various banking characteristics in the right ballpark.[17]

Costly bank default is calibrated based on the prior discussion in Section 2.3. I assume that default of a bank in the 90th percentile of the assets distribution in the model corresponds to a banking crisis in a developed economy as defined in Laeven and Valencia (2018) using the World Bank methodology. Similarly, default of a bank in the 10th percentile of the model distribution corresponds to a crisis in an emerging economy. I use a polynomial of order one to map each bank's size to the cost of default:

$$\text{Default Cost(j)} = d_1 + d_2 k(j) \tag{26}$$

where $d_1$ and $d_2$ are set to 0.0511 and 0.0075, respectively. These parameters help match the average output loss of 7.6% and the distribution of losses that range from 4.67% to 11.67% in the data.

## 4.2 Validation

**Moments** Table 2 lists all key banking moments in the data and in the model. Empirical data comes from the Call Reports. The sample covers the 2010:q1-2019:q4 period. Model data comes from a stochastic simulation of the baseline economy with N=1 intermediaries and T=2,000 quarters. Table 5 in the Online Appendix describes how we construct and define each variable or ratio. We start with the distribution of markups. The model does a good job of matching the average

and quantity elasticities with respect to instrumented shocks to local credit demand. The author finds that the average nationwide elasticity is 1.2, yielding a large average markup of 6.

[16] A noteworthy computational nuance is that a larger superelasticity increases the dispersion of the markup distribution but also imposes a tighter mechanical limit on bank assets. Bank-specific elasticity of demand may never drop below unity, meaning the limit on assets is $\theta^{\frac{\theta}{\epsilon}}$. We keep track of whether this limit binds on any point of the state space.

[17] $\rho_\xi$ is one of the key parameters in the model that directly impacts concentration in the banking sector. A large $\rho_\xi$ (e.g. 0.99) achieves a high degree of concentration and a very right-skewed distribution of leverage, which brings the model closer to the data. On the other hand, Galaasen et al. (2020) find that idiosyncratic borrower shocks that impact bank portfolio outcomes are volatile but not autocorrelated. We therefore set $\rho_\xi$ to 0.529 which is a compromise between empirical evidence on the persistence of idiosyncratic credit shocks and the ability of the model to match banking distributions perfectly.

Table 2: **Moments**

| | Data | | | Model | | |
|---|---|---|---|---|---|---|
| | Mean | 10% | 90% | Mean | 10% | 90% |
| Markups | 1.44 | 0.99 | 1.98 | 1.44 | 1.35 | 1.53 |
| Probability of Default | 1.07% | 0.87% | 1.37% | 1.02% | 0.00% | 2.97% |
| Real Cost of Default | 7.60% | 4.67% | 11.67% | 7.93% | 6.79% | 8.93% |
| Net Interest Income / Assets | 3.45% | 2.71% | 4.21% | 2.66% | 0.88% | 5.33% |
| Non-Interest Expenses / Assets | 2.96% | 1.98% | 4.00% | 2.21% | 1.80% | 2.58% |
| Interest Expenses / Assets | 0.62% | 0.22% | 1.15% | 0.61% | 0.45% | 0.82% |
| Book Leverage | 9.42 | 6.68 | 11.98 | 6.09 | 4.84 | 7.26 |

Notes: key moments in the model and in the data. Variables are defined in Table 5. Probabilities are annualized. Markups are absolute. Data source: Call Reports. Quarterly data is over 2010:q1-2019:q4.

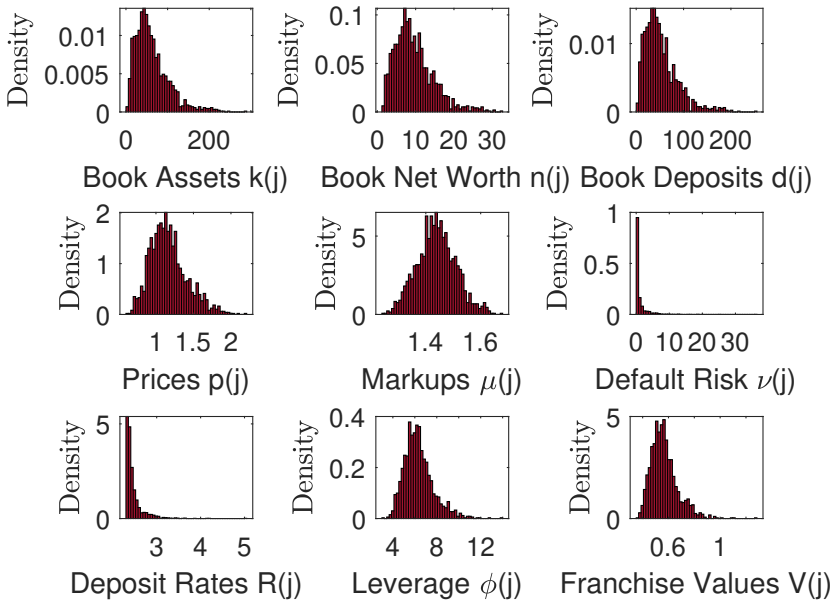
markup. The dispersion is slightly lower in the model, a point that is related to our discussion in Footnote 16. The success of nailing down the average markup is to a large extent due to flexibility of the Kimball aggregator.

The model also succeeds in getting exit risk right. The average probability of default (1.02%) is almost exactly equal to the empirical counterpart (1.07%). The dispersion of default risk is slightly greater than in the data. Real costs of default, including the mean and percentiles of the distribution, also match the data well. The net interest margin, non-interest expenses, and interest expenses ratios are all in line with the data as well.

Finally, bank leverage ratios are generally lower than in the data. The reason for this is the following mechanism. The presence of idiosyncratic bank return shocks creates a powerful precautionary lending motive for banks. Banks in the model are effectively risk averse, because the household is, and are thus rushing to outgrow the leverage constraint and the positive default risk region as soon as possible. This leads to a rapid accumulation of net worth. Interestingly, this implies that the riskier the economy is exogenously, the less risky it can become endogenously. This relationship arises in various setups, such as in Fostel and Geanakoplos (2008). Exogenous constraints on the precautionary lending motive, such as a lower bound on the deposit rate or additional lending adjustment costs could potentially help solve the issue and increase equilibrium leverage.

**Distributions** We now look at all stationary cross-sectional distributions that the model generates. Figure 7 plots univariate histograms for bank assets k(j), net worth n(j), deposits d(j), market leverage $\phi$(j) relative prices p(j), markups $\mu$(j), default risk $\nu$(j), deposit rates $\bar{R}$(j), and franchise values V(j). In line with the data, the credit market is concentrated, i.e. there is a small fraction of large and

Figure 7: **Stationary Distributions in the Model**




Notes: Model-generated stationary distributions.

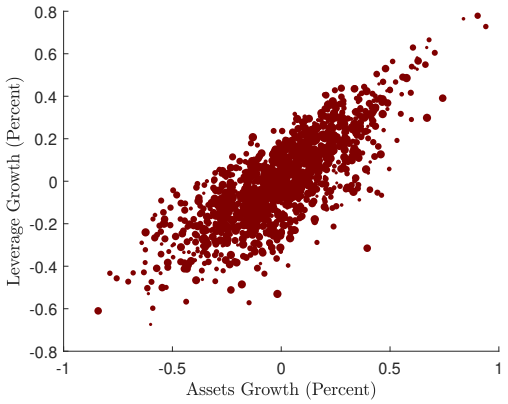
profitable intermediaries with a significant market share of assets, deposits, and net worth. The distribution of markups has the same dispersed and slightly right-skewed shape as in, for example, Pasqualini (2021): the right tail is driven by the largest banks in the economy who charge the highest markups. Distributions of default risk and deposit rates, which feed into relative prices through the marginal cost channel, are of a similar right-skewed and dispersed shape. Here, in contrast, the right tail is driven by the low net-worth intermediaries with risky balance sheets and high marginal costs.

**Cross-Sectional Correlations** We now focus on the three essential cross-sectional patterns from the data that the model reproduces. They are important because together they constitute the policy trade-offs which we define and discuss in Section 5. First, in line with the data, the model generates a positive cross-sectional correlation between bank leverage and assets growth. Figure 8 visualizes the result in Panel (a). The two variables come from a stochastic simulation of the model with N=1 intermediaries and T=2,000 quarters. The positive association can be clearly seen from the scatter plot.

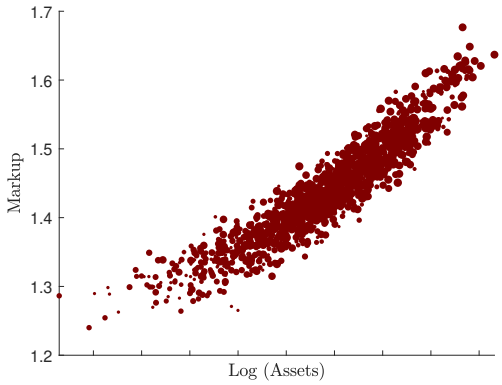
Second, the model correctly predicts that larger banks choose to charge higher markups $\mu$(j). This can be seen clearly from Panel (b) of Figure 8. Interestingly, the shape of the markup function is similar in both the model and the data - it is increasing and slightly convex. Third and finally, the

Figure 8: **Cross-Sectional Patterns in the Model**

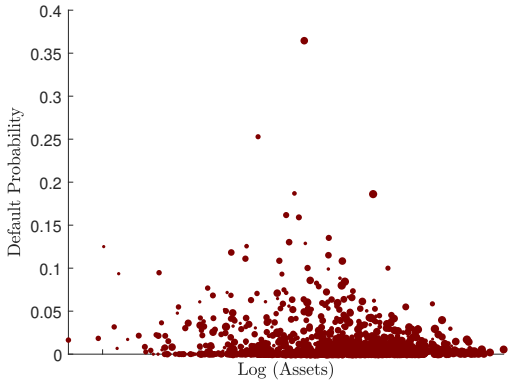
(a) Assets and Leverage




(b) Assets and Markups




(c) Assets and Default Risk

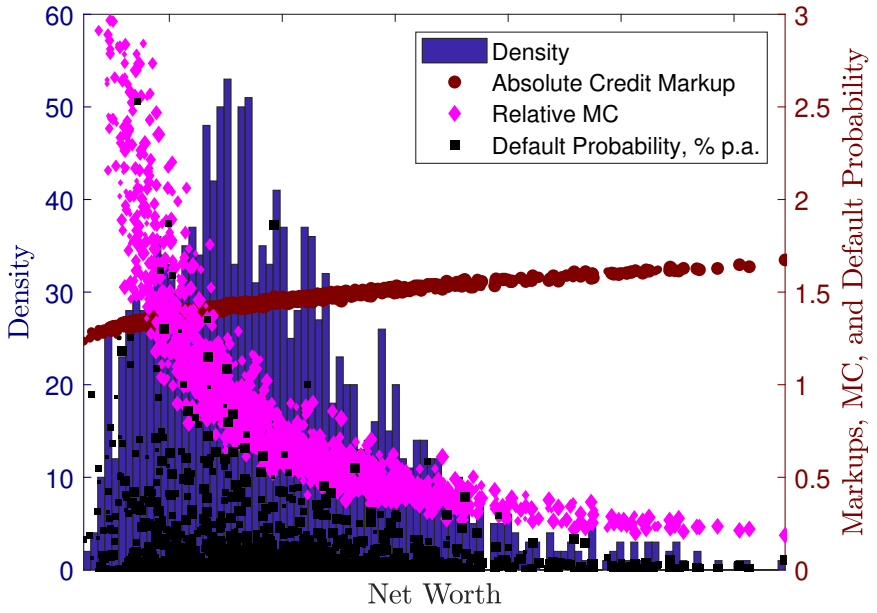



Notes: Cross-sectional relationships between measures of bank size, default risk, book leverage, and markups in the model. All panels show scatter plots based on a stochastic simulation of the baseline model with N=1 intermediaries and T=2,000 quarters.

model predicts that the cross section of default risk $\nu(j)$ is concentrated in the left tail of the assets distribution. Figure 8, Panel (c) shows the relationship on a scatter plot. It can be readily seen that large intermediaries are essentially risk-free. In the lower percentiles of the distribution, however, exit risk escalates rapidly and goes beyond 10%.

I conclude this section by emphasizing that although the model does not match every aspect of the U.S. banking data perfectly, it does very well in capturing the relationships between scale, competition, and stability - the three facets of the policy trade-off we are about to discuss next.

Figure 9: **The Banking Industry Trilemma**




Notes: distribution of bank net worth and scatter plots of markups, relative marginal costs, and probabilities of default.

# 5 Quantitative Analysis of Bank Regulation

In this section I first define the tradeoff between bank competition, stability, and efficiency. I then discuss how introducing heterogeneous capital requirements or deposit insurance affects the macroeconomy through the prism of this tradeoff. I proceed by solving for constrained-efficient allocations of a social planner. Finally, I analyze aggregate effects of the too-big-to-fail subsidy.[18]

## 5.1 The Competition, Efficiency, and Stability Trilemma

The tradeoff between financial competition, efficiency, and stability arises due to the interaction of three channels: economies of scale, endogenous competition, and financial stability. Figure 9 visualizes the mechanism. The picture plots the stationary distribution of bank net worth in the background. Overlayed are scatter plots for absolute markups $\mu(j)$, relative marginal costs, and absolute probabilities of default $\nu(j)$ in percent p.a. The *economies of scale* channel is represented by the negative relationship between marginal costs and net worth - larger banks are more cost-efficient and have a greater marginal propensity to lend MPL(j). The *endogenous competition* channel is seen from the positive relationship between markups and net worth. Finally, the *financial stability*

[18] If any of the results or claims are not clear from the figures, Table 3 summarizes allocations across all scenarios.

channel is seen from the negative relationship between the default probability and net worth. The trilemma exists because banks that are efficient and stable are the same ones that charge higher markups. Efficiency gains from having more banks with low $\nu(j)$ and high MPL(j) is counteracted by them contributing to a higher average markup and, as a result, greater welfare losses from bank market power.

At the heart of the trilemma are two intertwined *bilateral* tradeoffs. First, the canonical competition-stability tradeoff. Monopolistic competition allows high-net-worth banks to reach greater equilibrium franchise values through higher markups. This, in turn, reduces appetite for private risk-taking, lowering the probability of insolvency in equilibrium. Second, the competition-efficiency tradeoff. High-net-worth intermediaries charge higher markups but they are also more efficient from the cost- and productivity standpoints, as discussed previously.[19] Each of the two bilateral tradeoffs can be viewed to rely on the variable markups channel. However, even with constant markups we still entertain an efficiency-stability frontier, similarly to Ranciere et al. (2008). Macroeconomic aggregates will be negatively associated with systemic stability, in equilibrium, as we will see by the end of this section.

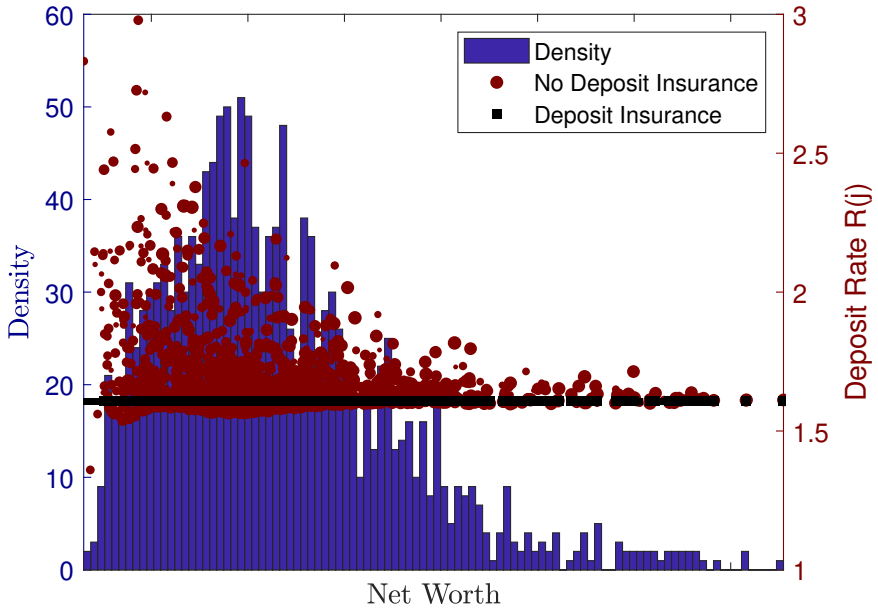
It is crucial to emphasize that the trilemma does not imply that it is *impossible* for the regulator to improve *net* welfare. Net quantitative effects always depend on calibration of the credit demand superelasticity, the cost of default function, and of the idiosyncratic risk process. Suppose we consider an economy where the largest intermediaries are state-run. Physical costs of default of banks in the right tail would therefore potentially always outweigh any efficiency losses from high markups in "normal times". We are merely claiming that if a regulator attempts to improve any side of the trilemma, one or all of the remaining two dimensions would necessarily deteriorate as a matter of unintended consequences. Whether the policy shock or the unintended side-effects dominate is a quantitative question.

## 5.2 Deposit Insurance

We now explore how in our calibrated model various policy schemes affect the macroeconomy through the prism of the aforementioned trilemma. We begin with deposit insurance. When deposit guarantees are switched on, the distribution of equilibrium deposit rates $\bar{R}(j)$ is flat. To achieve this as an endogenous result, banks continue to take as given their individual default probability $\nu(j)$, but we break the mapping between $\nu(j)$, the deposit recovery rates x(j), and rates $\bar{R}(j)$. We then

[19]There is an alternative view that argues larger banks are *not* more efficient than smaller banks (Huber, 2021). My model is consistent with Huber (2021) because of the endogenous competition channel. The author considered a quasi-natural experiment that shocks bank size without affecting local competition. In my framework, increases to bank size conditional on holding competition constant do necessarily improve efficiency. However, if markups are endogenous and heterogeneous, as they seem to be in reality, then unconditional increases to bank size have ambiguous effects on net efficiency if markups also rise. This is precisely the competition-efficiency trade-off.

Figure 10: **Deposit Insurance Scheme**



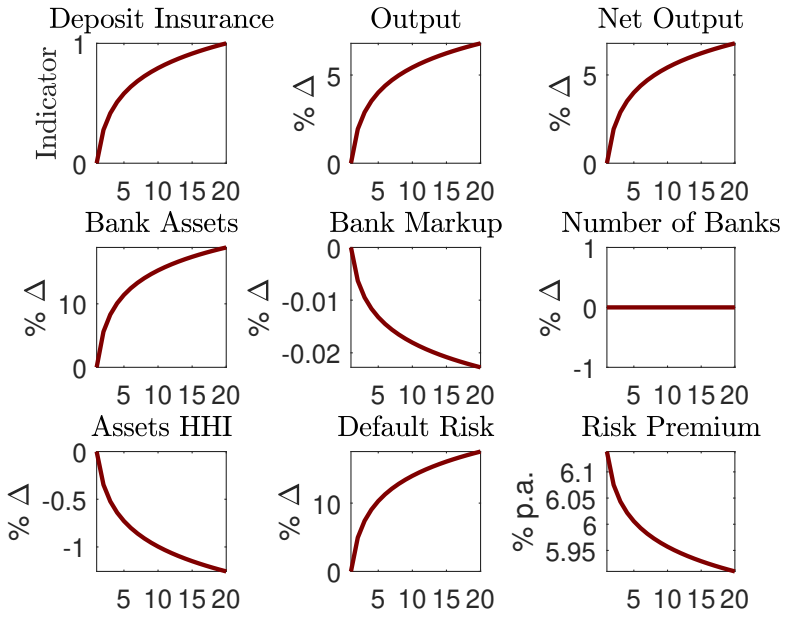

Notes: distribution of bank net worth and scatter plots for deposit rates in the economy with deposit insurance. Black squares represent equilibrium rates when guarantees are turned on and red circles the counterfactual rates if guarantees were turned off.

re-solve the model as usual. In other words, there is no equilibrium pass-through from balance sheet riskiness to the prices of debt. The government promises to honour any deposit shortfalls due to endogenous bank exit, and both banks and the household interpret this (rationally) as a unitary and inelastic recovery rate on all deposits in the distribution. We assume that the government funds the scheme via non-distortionary lump-sum taxes on the household.

Figure 10 shows the outcome of this policy. In the background is the new stationary distribution of bank net worth which is consistent with the equilibrium with deposit insurance. The flat-lined black scatter plot shows the invariance of the equilibrium $\bar{R}(j)$ with respect to bank size. For contrast, the red scatter plot represents the counterfactual rates if guarantees were turned off; notice the usual inverse relationship with n(j) in that case. It can be seen from the Figure that the biggest beneficiaries from the introduction of deposit guarantees are low-net-worth banks with high marginal costs.

Figure 11 demonstrates the macroeconomic effects of deposit insurance. At t=0, we start from the baseline stationary industry market equilibrium. At t=20, the economy has permanently converged to its version with full deposit insurance. Lending, output, and net output (net of realized costs of bank default) all increase since funds are now cheaper to obtain. In line with much of the theoretical and empirical evidence on the interactions between risk-taking and deposit insurance,

Figure 11: **Macroeconomic Effects of Deposit Insurance**




Notes: macroeconomic effects of switching on deposit guarantees. Net output is defined as output $Y_t$ net of real costs of bank default.

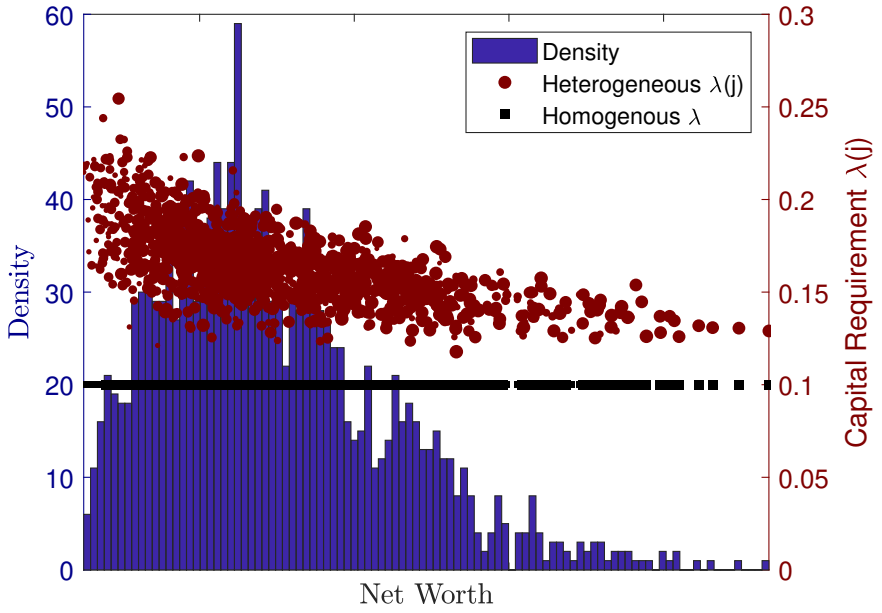
we see a positive effect on average default risk. Since the deposit insurance scheme favors small banks by more, the market share of of low-net-worth banks increases and concentration (Assets HHI) falls. As a result, the average markup falls. Finally, aggregate risk premium, defined as $R^k - \bar{R}$ (annualized return on aggregate capital minus the average deposit rate) falls since aggregate quantities rise and prices fall - both effects lower $R^k$.

Notice how the mechanism of the trilemma holds - introducing deposit insurance has increased lending and output but also raised systemic riskiness. Quantitatively, net output has still grown because default costs turn out to be negligble. Recall that the number of banks is time invariant because for now we still operate with an exogenous entry margin.

## 5.3 Heterogeneous Leverage Regulation

The second market intervention that we consider is heterogeneous capital requirements. One possibility to impact private risk-taking and aggregate fragility is to impose regulatory limits on $\phi(j)$. In practice, this corresponds to micro-prudential regulation which is a common practice by governments around the world. Recall that in the market economy, leverage falls with bank size while $\lambda$ is homogenous across all banks. We now consider a scenario where $\lambda(j)$ is ex-ante heterogeneous and falls linearly with bank net worth. Banks in the top decile of the distribution

Figure 12: **Heterogeneous Capital Requirements**



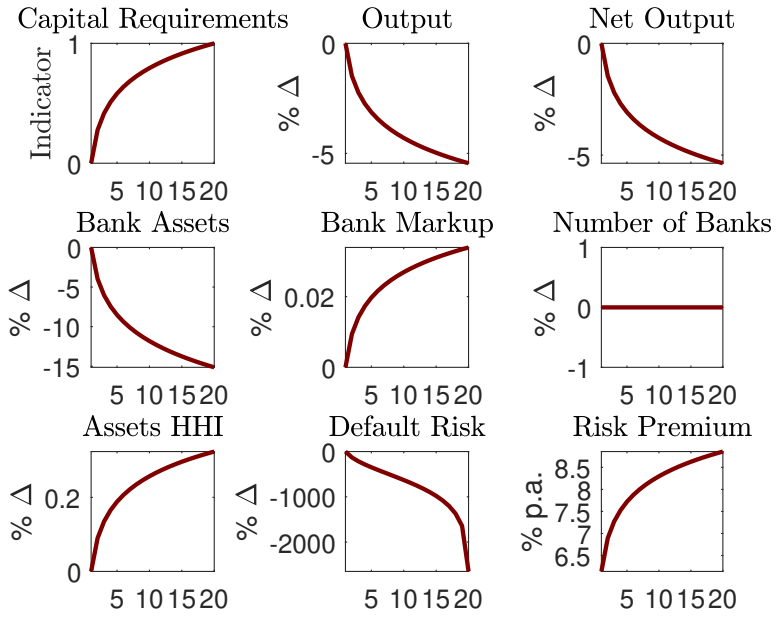

Notes: distribution of bank net worth and scatter plots for $\lambda$ under homogenous and heterogeneous capital requirement regimes.

face the same $\lambda_{\iota=10} = 0.1$ as before. However, banks in the lowest decile face $\lambda_{\iota=1} = 0.3$. All banks in the deciles that are in between face a $\lambda(j)$ that is interpolated based on the exogenous grid of net worth and their position in the distribution. What this policy is designed to achieve is to restrict leverage of precisely those intermediaries who are the most likely to have high leverage to begin with.

Figure 12 shows how the policy works in the model. Overlayed on the new equilibrium distribution of net worth are the homogenous $\lambda$ from the baseline economy and $\lambda(j)$ from the economy with capital requirements. The negative slope of the $\lambda(j)$ scatter plot implies that the policy has achieved its desired objective - limitations on leverage are proportional to actual leverage, here summarized by n(j) as the sufficient summary statistic. Recall that $\phi(j)$ falls with n(j), as demonstrated and discussed earlier in Figure 26.

Figure 13 portrays the macroeconomic effects of this policy. Similarly to before, t=0 and t=20 correspond to the baseline regime and the case with heterogeneous $\lambda(j)$, respectively. We see that all aggregate quantities have fallen, including lending, output, and net output. This is driven by the fact that the leverage constraint is now tighter precisely for the agents for which it is more likely to bind - the low-net-worth banks. As a result, aggregate demand for deposits and bank leverage are down. Notice how default risk has fallen by a considerable amount - the average

Figure 13: **Macroeconomic Effects of Heterogeneous Capital Requirements**



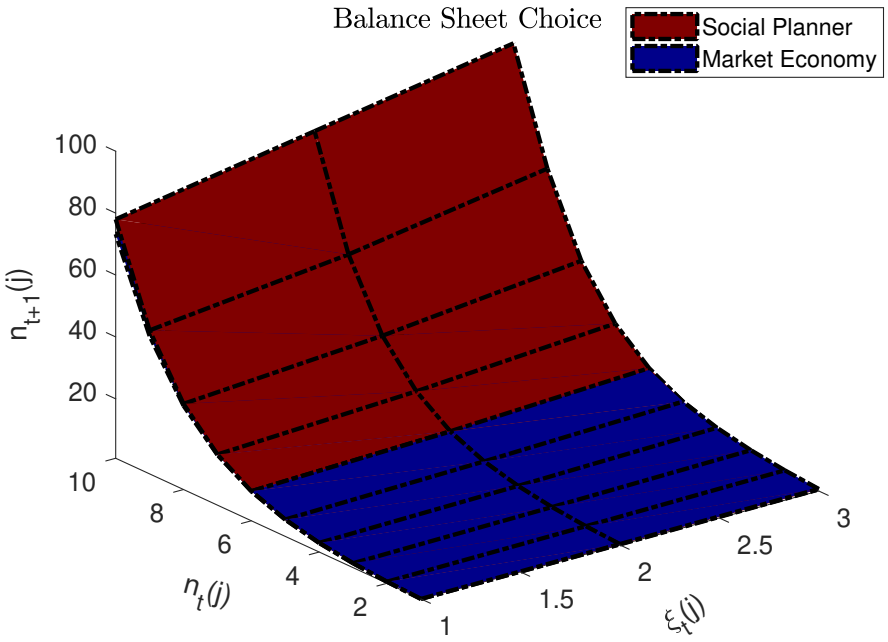

Notes: macroeconomic effects of switching on heterogeneous capital requirements.

probability of default is almost 0%. The policy is highly successful in terms of reducing systemic financial fragility. The tradeoff here comes from the reduction in efficiency, intermediation, and production. Risk premia are up because aggregate capital is down and prices are (slightly) up. Concentration is slightly up because the new $\lambda(j)$ regime is detrimental for precautionary lending growth for practically all banks in the economy, with the exception of the largest intermediaries who had outgrown the constraint completely. Because the distribution is more concentrated, the average markup increases. All in all, reduction in systemic risk is achieved at the cost of higher aggregate markups and lower efficiency.

## 5.4 Constrained Efficiency and Optimal Policy

**Constrained efficiency** We now consider constrained-efficient allocations of a hypothetical social planner as a stepping stone for optimal policy analysis. The planning problem is identical to that of the baseline economy with one crucial exception. In this section, the planner picks the quadruple $\{k, d, p, \mu\}$ in order to maximize the franchise value V while understanding that $R^T$ is *endogenous* through the impact of the quadruple on $R^k$ and P. Consider the law of motion of net worth that the social planner faces:

Figure 14: **Market Equilibrium and Constrained Efficient Allocations**




Notes: Market-based and social planner's allocations from the stationary equilibrium.

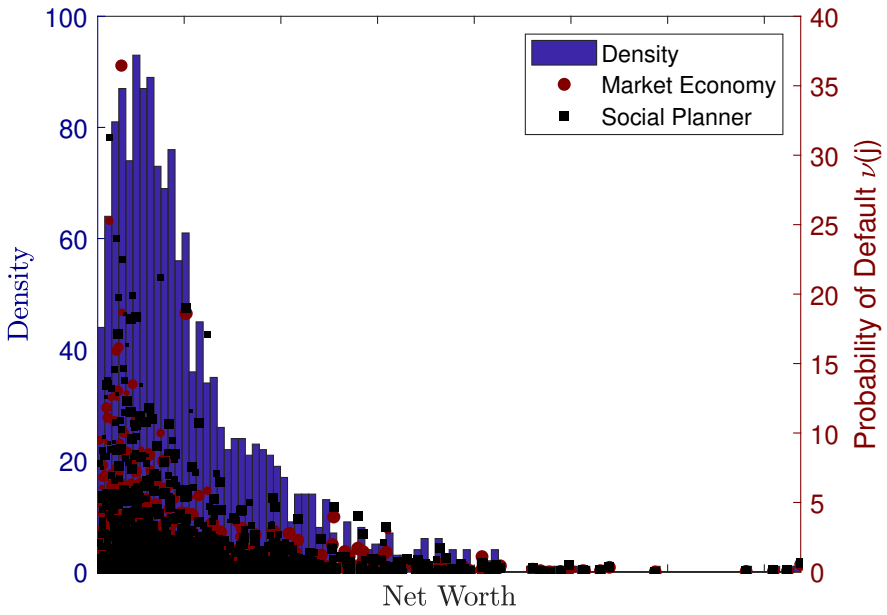
$$n_{t+1}(j) = R^T\Big(n(j), \xi(j), \{k_t(j), d_t(j), p_t(j)\}\Big)p_t(j)k_t(j) - \bar{R}_t(j)d_t(j) - \frac{1}{\zeta_1}k_t(j)^{\zeta_2} \tag{27}$$

Compare this formula to Equation 10 from the market 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 Section F of the Online Appendix for more details on the numerical algorithm.

Figure 14 presents the two-dimensional optimal choice of next-period bank net worth $n'(j)$. We contrast decisions of the social planner with the market outcome. Comparing the two cases reveals that misallocation is present in the decentralized equilibrium along both the net worth and idiosyncratic risk dimensions. Specifically, the market outcome yields too *little* lending because of an aggregate credit supply externality.[20] Monopolistic credit competition leads to underutilization of risky capital as a resource in production. Unlike the social planner, individual banks do not internalize the impact of their private choices on aggregate returns. In addition, misallocation is more severe for higher levels of net worth. This is consistent with the idea that markups are variable

[20]This is a credit market version of the classical aggregate demand externality (Blanchard and Kiyotaki, 1987; Farhi and Werning, 2016).

Figure 15: **Systemic Risk Implications of Constrained Efficiency**



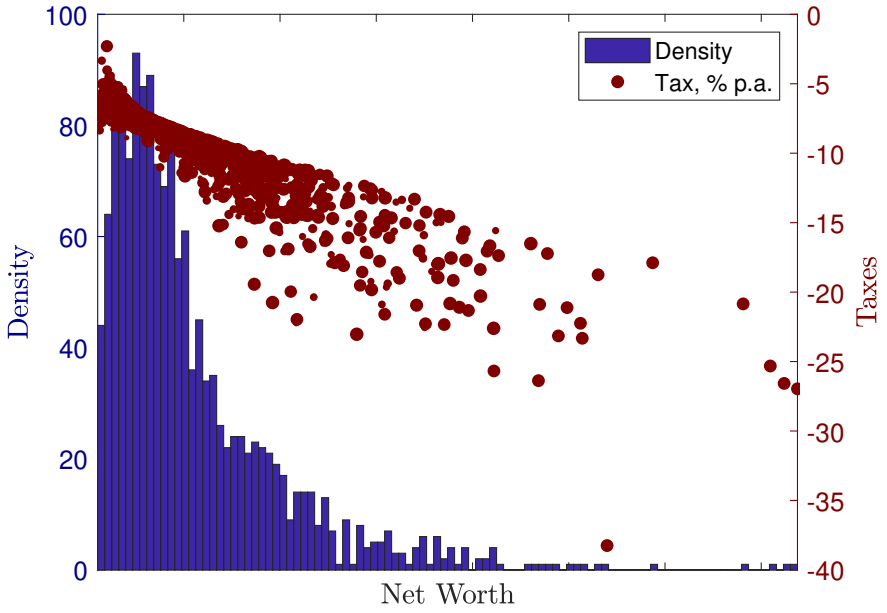

Notes: distribution of bank net worth and scatter plots for $\nu(j)$ under alternative market regimes.

and increase with net worth. Recall that demand for firm claims is more satiated in the right tail of the bank size distribution.

Figure 15 shows how equilibrium financial stability responds to social planner's actions. We continue to define financial stability as the probability of bank default due to insolvency $\nu(j)$. We plot the new stationary distribution of net worth and $\nu(j)$ scatter plots that correspond to the constrained efficient (black square) and market (red circle) allocations. Here we observe that the social planner's solution induces an increase in system-wide default risk. Low-n(j) intermediaries become particularly more risky. This result is a case in point of the financial stability-competition trade-off (Hellman et al., 2000). Specifically, the social planner targets the credit supply externality by reallocating resources towards agents with the highest marginal propensity to lend - the bigger banks. However, smaller intermediaries are fundamentally more prone to insolvency risk to begin with. Small, risky banks become relatively riskier. Average probability of default, as a result, goes up and the economy is more fragile.

**Optimal policy** We decentralize constrained efficient allocations with taxes on banks gross returns. Importantly, these policies are size- and income-dependent because misallocation and markups correlate with the joint distribution of bank net worth and idiosyncratic risk. Theoretically, gross returns taxes are easier to operationalize because they target specifically the wedge in the

Figure 16: **Optimal Policy**




Notes: distribution of bank net worth and scatter plots for the optimal $\tau(j)$, in percent p.a.

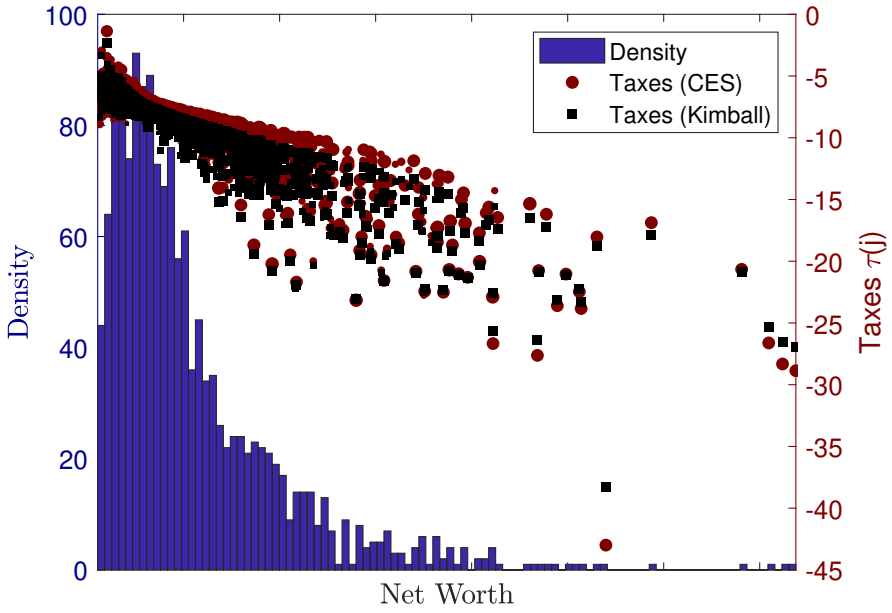
bank-specific total portfolio return process and the law of motion of net worth. Specifically, we conjecture a size and idiosyncratic return specific tax rule $\tau\Big(n(j), \xi(j)\Big)$ and impose it on the market equilibrium. Computationally, we assume that taxes are polynomials in $n(j)$ and $\xi(j)$ and solve for coefficients that are consistent with a minimal distance between the equilibrium and the social planner allocations. 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 bank net worth with tax policies is now:

$$n_{t+1}(j) = R_t^T(j)\left[1 - \tau\Big(n(j), \xi(j)\Big)\right] p_t(j)k_t(j) - \bar{R}_t(j)d_t(j) - \frac{1}{\zeta_1}k_t(j)^{\zeta_2} \tag{28}$$

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

Figure 16 plots the stationary distribution of net worth from the social planner's problem with the scatter plot for optimal taxes $\tau(j)$. Notice how all intermediaries in the state space receive a *subsidy*. The subsidy is the highest (in absolute terms) for *big* banks. The intuition for this result is related to the economies of scale channel: marginal propensity to lend (MPL) increases with bank net worth. The social planner finds it most efficient to correct the under-lending externality by stimulating/subsidizing lending of those with the highest marginal propensity to respond to

Figure 17: **Optimal Taxes under CES and Kimball Aggregators**

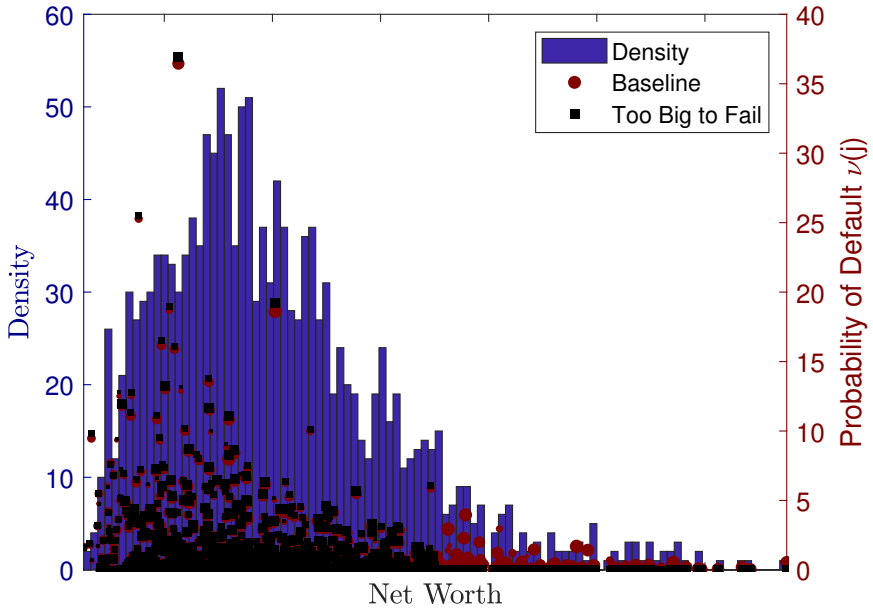



Notes: optimal taxes $\tau$(j) under constant (CES) and variable (Kimball) markups.

taxes. In general equilibrium, this increases aggregate output and household consumption. In the stationary distribution, the annualized tax ranges from -38% to -2% with the average tax of about -9.38% per year.

**The Role of the Aggregator** An interesting auxiliary exercise is to compare normative implications under the two alternative regimes for bank markups: constant and variable. Figure 17 plots the scatter plot for optimal bank taxes under the Kimball and CES aggregators. Average taxes for the CES and Kimball economies are -8.75% and -9.38%, respectively. In the Kimball economy markups are not only dispersed relative to CES but also slightly right-skewed. The average markup is thus slightly higher because the size distribution is also right-skewed. In both scenarios, subsidies are also heavily size-dependent: they increase with net worth, mirroring the shape of the MPL function and the economies of scale channel. Underutilization of capital is thus greater in the Kimball economy, the wedge between the constrained best and the market outcome is greater, and the corrective tax (subsidy) that the planner wishes to impose is higher. The CES aggregator therefore potentially "understates" the welfare costs of lending market power of banks.

Figure 18: **Too Big to Fail**




Notes: distribution of bank net worth and scatter plots for $\nu(j)$ for the baseline economy with and without the TBTF subsidy.

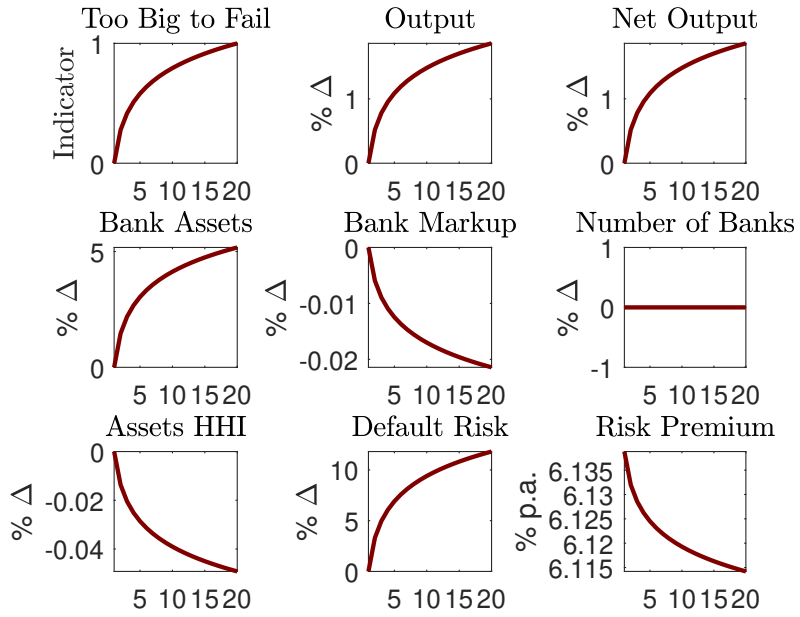
## 5.5 Too Big to Fail

Absence of effective bank failure-resolution rules and laws pre-Lehman meant that systemically important banks, particularly those with U.S. headquarters, benefited from implicit "too-big-to-fail" (TBTF) subsidies. Probability of an ex-post government bailout of large financial institutions was close to one, which was priced by the market into lower debt financing costs after the adjustment for insolvency and illiquidity risks. Conditional on this safety net being part of the environment, market participants lose their incentive to monitor the intermediaries and banks lose the incentive to act prudently, which further exacerbates the problem (Stern and Feldman, 2004).[21]

In the context of our model, we operationalize the TBTF hazard problem the following way. The probability of default $\nu(j)$ of any bank in the top decile of the distribution of net worth is zero, regardless of balance sheet properties. We pick the top decile simply for quantitative tractability. The bottom nine deciles instead face a $\nu(j)$ that is consistent with their size-risk profile as usual. The policy function for $\nu(j)$ is therefore kinked, and all banks understand this. The pass-through from $\nu(j)$ to $\bar{R}(j)$ functions normally - all banks face the cost of debt that is consistent with their probability of default. This implies that some banks will face an exogenously imposed "cost of

[21] In recent years, multiple studies have found that the TBTF problem has declined after the passage of the Dodd-Frank Act. (Haldane, 2010; Atkeson et al., 2018)

Figure 19: **Macroeconomic Effects of Too Big to Fail**




Notes: macroeconomic effects of switching on the TBTF subsidy.

funds subsidy" which switches on only if the bank reaches a certain size threshold.

Figure 18 shows how the mechanism works. Observe how the scatter plot for $\nu(j)$ in the TBTF case is clearly kinked: banks in the right tail of the distribution face no default risk exogenously. In contrast, in the baseline economy some of the same banks in the right tail face a positive $\nu(j)$. In addition, it is interesting that most banks in the bottom nine deciles now face a *higher* $\nu(j)$. The TBTF subsidy, even if it affects only the largest intermediaries, makes leverage choices of all banks strategic complements (Farhi and Tirole, 2017). The subsidy reinforces the already strong precautionary lending motive - banks choose higher equilibrium leverage because it allows them to accumulate more net worth with less downside risk.

Figure 19 presents the macroeconomic effects of the TBTF problem. As usual, we are considering two regimes with an instantaneous transition. We see that the TBTF hazard raises output, net output, and financial intermediation activity. The subsidy allows large intermediaries to lever up by more, thus the positive effect on aggregate demand. Higher production comes at a cost of greater systemic riskiness. The positive impact on aggregate default risk is the result of strategic complementarity in risk-taking - the economy is more efficient but far less stable. This straightforward exercise illustrates how the TBTF subsidy may have caused a build-up of excessive financial risk prior to the Great Recession.

Table 3: **Summary of All Allocations**

| | Equilibrium | Constrained Efficiency | Deposit Insurance | Capital Regulation | Too Big to Fail |
|---|---|---|---|---|---|
| Output | 4.424 | 4.540 | 4.735 | 4.189 | 4.507 |
| Net Output | 4.420 | 4.535 | 4.730 | 4.189 | 4.503 |
| Book Leverage | 6.340 | 6.332 | 6.390 | 4.502 | 6.361 |
| Default Risk | 1.016% | 1.482% | 1.061% | 0.001% | 1.046% |
| Aggregate Markup | 1.437 | -0.345 | -0.023 | 0.034 | -0.021 |
| Risk Premium | 6.139 | 4.733 | 5.910 | 8.858 | 6.114 |

Notes: Macroeconomic and financial aggregates across different regulatory and market regimes. Markups are in % deviations relative to the equilibrium markup.

## 5.6 Summary of All Allocations

To summarize our findings across different regulatory regimes and market structures, we report all key aggregates in Table 3. We focus on output, net output, bank leverage, systemic risk, the aggregate markup, and the risk premium. Output is aggregate production from the stationary steady state. Net output is gross output adjusted for the real costs of realized bank default. Book leverage is the unweighted average of $\frac{k(j)}{n(j)}$. Systemic risk is defined as the average probability of bank default $\nu(j)$, annualized. The aggregate markup is the unweighted average of $\mu(j)$ for the market economy; for all other cases markups are represented as percentage deviations relative to the market economy. Risk premium is defined as $R^k - \bar{R}$, i.e. return on the risky asset minus the average interest rate on deposits, both annualized.

We start with the first column - the market economy - which is our benchmark for comparisons. Relative to the market equilibrium, constrained efficiency achieves a higher level of output but also leads to the highest probability of bank default among all the cases that I considered. Introduction of deposit insurance raises aggregate output at the cost of greater leverage and systemic risk. Heterogeneous capital requirements, on the other hand, virtually eliminate financial instability but reduce aggregate efficiency, raise markups, and increase the intermediation spread (risk premium). The too-big-to-fail externality increases both aggregate production and systemic riskiness.

Overall, we have seen that qualitatively no change to market structure or regulatory regime simultaneously improves output, stability, and competition. Quantitatively, conditional on our calibration strategy, the best outcome in terms of net output is the economy with deposit insurance. It is worth re-emphasizing that the argument of this paper is that the banking industry trilemma would always persist qualitatively. Of course, different parameterization approaches could amplify one arm of the trilemma relative to the others. For example, a calibration based on an emerging

economy with volatile financial markets could reverse the quantitative pecking order of policies, but not the qualitative prediction that the trade-offs are always there.

### 5.7 Additional Results and Applications

The Online Appendix provides further results and quantitative applications. In Section A I analyze applications of the model to the rise of banking concentration, emergence of fintech-intermediated credit, and intermediary asset pricing. Section B studies targeted bank-level stabilization policies such as equity injections and liquidity facilities. Section C simulates MIT shocks to aggregate productivity.

## 6 Conclusion

In this article I develop a novel macroeconomic framework for positive and normative analysis of macroeconomic transmission through bank heterogeneity. The model introduces two workhorse approaches in modern macro-finance - uninsurable idiosyncratic risk and imperfect competition - into the banking sector of the workhorse Gertler and Kiyotaki (2010) macroeconomic environment. The model is validated by replicating key cross-sectional patterns in the U.S. banking data. Bank heterogeneity matters for the design of various economic policies that run through the bank lending and market power channels. I analyze different regulatory schemes and issues such as deposit guarantees, heterogeneous capital requirements and the too-big-to-fail implicit subsidy. I also study optimal policy in a fully constrained-efficient version of the economy. Policy analysis at all levelts points at a *trilemma* for bank regulation. There is a trilateral trade-off between financial competition, stability, and efficiency. Through the lenses of this trilemma, I characterize auxiliary predictions of the framework for the rise of banking concentration, emergence of fintech-intermediated credit, unconventional targeted fiscal and monetary policy interventions, and intermediary asset pricing.

My model is tractable and can be readily extended to include additional parts.[22] An open-economy extension could be introduced, allowing us to study endogenous global financial cycles that are driven by heterogeneous, imperfectly competitive intermediaries. An extension with nominal rigidities could uncover a powerful channel of transmission that runs through bank heterogeneity in risk-taking and market power.

[22] Jamilov and Monacelli (2020) build on a variant of my framework with constant markups and introduce aggregate uncertainty. They study novel channels of business cycle amplification that arise from dynamic bank heterogeneity.

# 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.

Atkeson, A. G., A. d'Avernas, A. L. Eisfeldt, and P.-O. Weill (2018): "Government Guarantees and the Valuation of American Banks," .

Baqaee, D., E. Farhi, and K. Sangani (2021): "The Supply-Side Effects of Monetary Policy," *NBER Working Paper*, 28345.

Beck, T., A. Demirguc-Kunt, and R. Levine (2006): "Bank concentration, competition, and crises: First results," *Journal of Banking  Finance*, 30, 1581–1603.

Begenau, J., S. Bigio, J. Majerovitz, and M. Vieyra (2020): "A Q-Theory of Banks," *Manuscript*.

Begenau, J. and T. Landvoigt (2020): "Financial Regulation in a Quantitative Model of the Modern Banking System," *Working Paper*.

Benetton, M. (2021): "Leverage Regulation and Market Structure: A Structural Model of the UK Mortgage Market," *Journal of Finance*.

Benhabib, B., A. Bisin, and M. Luo (2019): "Wealth distribution and social mobility in the US: A quantitative approach," *American Economic Review*, 109.

Berger, A. N. and T. H. Hannan (1998): "The Efficiency Cost of Market Power in the Banking Industry: A Test of the "Quiet Life" and Related Hypotheses," *The Review of Economics and Statistics*, 80.

Berger, A. N. and L. J. Mester (1997): "Inside the black box: What explains differences in the efficiencies of financial institutions?" *Journal of Banking  Finance*, 21.

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

Bianchi, J. and S. Bigio (2020): "Banks, Liquidity Management and Monetary Policy," *Manuscript*.

Bigio, S. and Y. Sannikov (2021): "A Model of Credit, Money, Interest, and Prices," *NBER Working Paper*, 28540.

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

Boissay, F., F. Collard, and F. Smets (2016): "Booms and Banking Crises," *Journal of Political*

*Economy*, 124(2).

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," *Review 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.

Claessens, S., J. Frost, G. Turner, and F. Zhu (2018): "Fintech credit markets around the world: size, drivers and policy issues," *BIS Quarterly Review*.

Coimbra, N. and H. Rey (2019): "Financial Cycles with Heterogeneous Intermediaries," *NBER Working Paper*, 23245.

Constancio, V. (2016): "Challenges for the European Banking Industry," *Conference on "European Banking Industry: what's next?"*.

Corbae, D. and P. D'Erasmo (2020a): "Capital Requirements in a Quantitative Model of Banking Industry Dynamics," *NBER Working Paper*, 25424.

——— (2020b): "Rising bank concentration," *Journal of Economic Dynamics and Control*, 115.

Corbae, D. and R. Levine (2018): "Competition, Stability, and Efficiency in Financial Markets," *Jackson Hole Symposium: Changing market Structure and Implications for Monetary Policy*.

De Fiore, F. and H. Uhlig (2011): "Bank Finance versus Bond Finance," *Journal of Money, Credit and Banking*, 43.

De Loecker, J., J. Eeckhout, and G. Unger (2020): "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, Forthcoming.

Dempsey, K. (2020): "Capital Requirements with Non-Bank Finance," *Working Paper*.

Diamond, D. (1984): "Financial Intermediation and Delegated Monitoring," *Review of Economic Studies*, 51(3).

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).

Fernandez-Villaverde, J., S. Hurtado, and G. Nuno (2019): "Financial Frictions and the

Wealth Distribution," *NBER Working Paper 26302*.

Fostel, A. and J. Geanakoplos (2008): "Leverage Cycles and the Anxious Economy," *American Economic Review*, 98.

Gabaix, X. (2011): "The Granular Origins of Aggregate Fluctuations," *Econometrica*, 79(3).

Galaasen, S., R. Jamilov, R. Juelsrud, and H. Rey (2020): "Granular Credit Risk," *NBER Working Paper 27994*.

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.

Gromb, D. and D. Vayanos (2002): "Equilibrium and welfare in markets with financially constrained arbitrageurs," *Journal of Financial Economics*, 66.

Haldane, A. (2010): "The $100 Billion Question. Commentary," *Bank of England*.

He, Z. and A. Krishnamurthy (2013): "Intermediary Asset Pricing," *Journal of Financial Economics*, 103(2), 732–770.

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).

Huber, K. (2021): "Are Bigger Banks Better? Firm-Level Evidence from Germany," *Journal of Political Economy*, Forthcoming.

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. (2020): "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.

KEELEY, M. C. (1990): "Deposit Insurance, Risk, and Market Power in Banking," *The American Economic Review*, 80.

KIMBALL, M. (1995): "The quantitative analytics of the basic neomonetarist model," *Journal of Money, Credit and Banking*, 27(4).

KLENOW, P. J. AND J. L. WILLIS (2016): "Real Rigidities and Nominal Price Changes," *Economica*, 83.

LAEVEN, L. AND F. VALENCIA (2018): "Systemic Banking Crises Database: An Update," *IMF Working Paper*, 18/208).

LEE, S., R. LUETTICKE, AND M. RAVN (2020): "Financial Frictions: Macro vs Micro Volatility," *CEPR DP*, 15133.

MALIAR, L., S. MALIAR, AND F. VALLI (2010): "Solving the incomplete markets model with aggregate uncertainty using the Krusell-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.

MIDRIGAN, V., C. EDMOND, AND D. XU (2018): "How Costly are Markups," *NBER Working Paper*, 24800.

NUNO, G. AND C. THOMAS (2017): "Bank Leverage Cycles," *American Economic Journal: Macroeconomics*, 9(2).

PASQUALINI, A. (2021): "Markups, Markdowns and Bankruptcy in the Banking Industry," *Working Paper*.

RANCIERE, R., A. TORNELL, AND F. WESTERMANN (2008): "Systemic Crises and Growth," *Quarterly Journal of Economics*, 123(1).

REINHART, C. M. AND K. S. ROGOFF (2009): "The Aftermath of Financial Crises," *American Economic Review*, 99.

REPULLO, R. (2004): "Capital requirements, market power, and risk-taking in banking," *Journal of Financial Intermediation*, 13, 156–182.

RIOS RULL, V., T. TAKAMURA, AND Y. TERAJIMA (2020): "Banking Dynamics, Market Discipline and Capital Regulations," *Manuscript*.

ROMER, C. D. AND D. H. ROMER (2017): "New Evidence on the Aftermath of Financial Crises in

Advanced Countries," *American Economic Review*, 107.

SCHULARICK, M. AND A. M. TAYLOR (2012): "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-2008," *American Economic Review*, 102.

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

STERN, G. AND R. FELDMAN (2004): "Too Big to Fail: The Hazards of Bank Bailout," *Brookings Institution Press*.

UHLIG, H. (2010): "A model of a systemic bank run," *Journal of Monetary Economics*, 57.

WANG, Y., T. WHITED, Y. WU, AND K. XIAO (2020): "Bank Market Power and Monetary Policy Transmission: Evidence from a Structural Estimation," *NBER Working Paper*, 27258.

WHEELOCK, D. C. AND P. WILSON (2012): "Do Large Banks Have Lower Costs? New Estimates of Returns to Scale for U.S. Banks," *Journal of Money, Credit and Banking*, 44.

WHEELOCK, D. C. AND P. W. WILSON (2018): "The evolution of scale economies in US banking," *Journal of Applied Econometrics*, 33.

# Online Appendix for "A Macroeconomic Model with Heterogeneous Banks"

Rustam Jamilov

London Business School

April 8, 2021

## Contents

# A Quantitative Applications

In this section I explore various applications of the framework and of the bank policy trilemma. First, we study predictions of the model for the ongoing global rise in banking concentration. We proceed by examining implications of the emergence of fintech-intermediated credit. Finally, we conclude with implications for intermediary asset pricing models and empirics.

## A.1 The Rise of Banking Concentration

The banking industry around the world is becoming more and more concentrated (Corbae and D'Erasmo, 2020b; Constancio, 2016). We do not take a stance on the *cause* behind the rise of concentration. Instead, we quantify the impact of various distributions of banks on the macroeconomy by fitting several counterfactual cross-sectional distributions of bank assets into the stationary general equilibrium and re-evaluating all endogenous variables that would be consistent with them. Counterfactual distributions are generated exogenously by drawing sequences of bank assets $k_1 \dots k_N$ from well-known continuous probability densities such as Uniform or Pareto. We fit each generated sequence into the model, re-compute all policy functions, but do not run the step which calculates new distributions. In other words, we solve for partial-equilibrium policy functions that are consistent with the exogenously constructed distributions.

We consider 3 broad families of densities: Uniform, Lognormal, and Pareto. For the uniform density, we generate N=2,000 random numbers from the interval $[0.5K_{ss}, 1.5K_{ss}]$, where $K_{ss}$ stands for the level of aggregate capital in the market equilibrium. For the lognormal density, we draw from $P(\mu_k, \sigma_k^2)$, where $\mu_k$ and $\sigma_k$ are, respectively, the mean and standard deviation of the k(j) distribution from the stationary equilibrium. For the Pareto density, we follow Gabaix (2009) and consider the Pareto I family with a power parameter of $\alpha = 2$.[1]

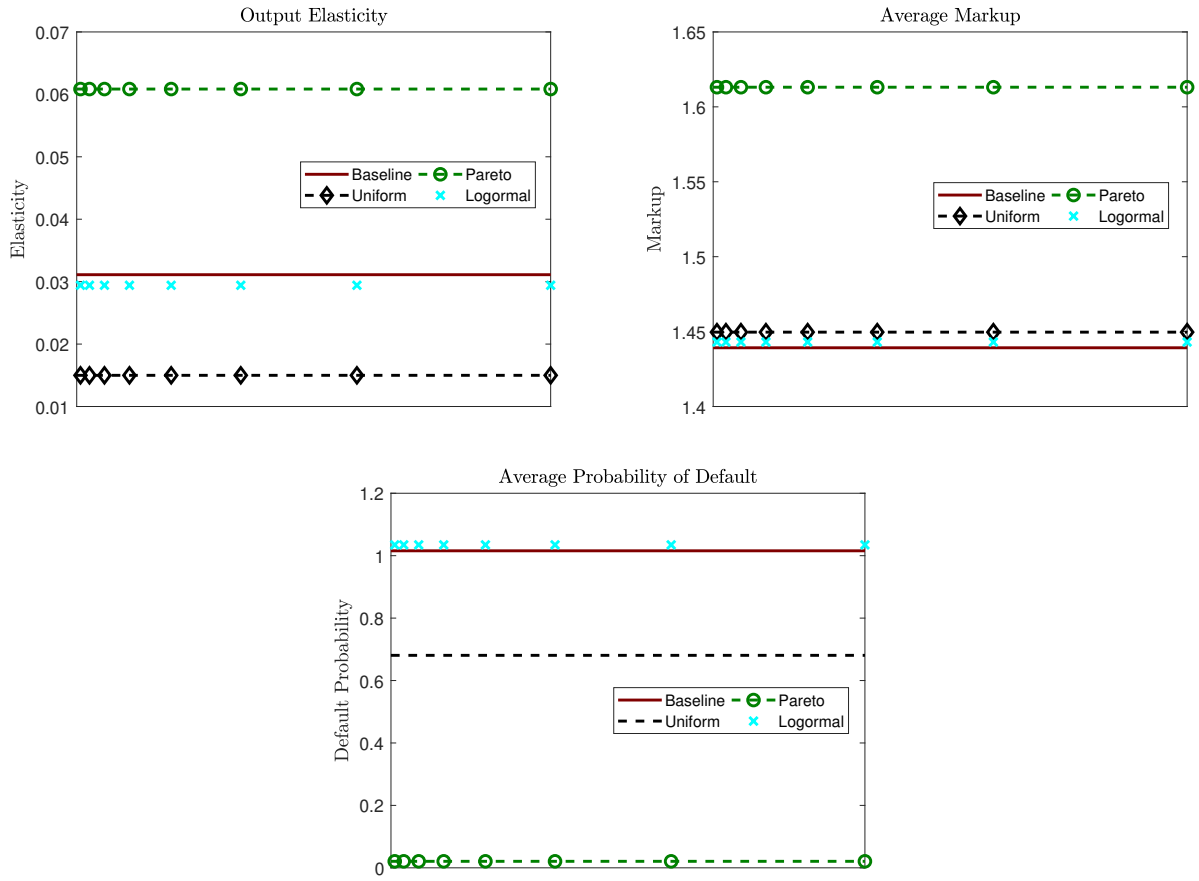
We focus on three aggregate variables of interest - the output elasticity of uniform bank net worth shocks, the average markup, and the average probability of default. These three objects summarize the three dimensions of the banking industry trilemma. The output elasticity can be defined with the help of the previously analyzed marginal propensities to lend (MPL):

$$\frac{\partial Y}{\partial N} = \underbrace{\frac{\partial Y}{\partial K}}_{\text{MPK}} \times \int_{\mathbf{B}} \underbrace{\frac{\partial k(j)}{\partial n(j)} \mu(dn, d\xi)}_{\text{MPL(j)}} \tag{29}$$

where the first term on the right-hand side is the marginal product of capital and the second term is the aggregate MPL. We treat a high output elasticity as a symptom of high efficiency in the lending market.

[1] The scale parameter is chosen to be a factor of $k_{min}$, i.e. the minimum level of assets from the market economy.

Figure 20: **Macroeconomic Effects of Alternative Banking Distributions**




Notes: Output elasticities, probabilties of default, and aggregate markups across alternative cross-sectional distributions.

Figure 20 presents the results of this exercise. We observe that the output elasticity with respect to uniform net worth shocks is highest for the Pareto economy, followed by the baseline, lognormal, and uniform economies. Intuitively, the degree of right-skewness can be viewed as a sufficient statistic for the elasticity, and thus efficiency. Similarly to what we concluded based on Figure 9, the bigger the share of high-n(j) banks the higher aggregate efficiency gets. Because of the economies of scale channel, larger banks have both a greater MPL and a lower marginal propensity to price (MPP). The fact that the Pareto economy, which is more concentrated than our baseline model, has a higher elasticity is proof of the mechanism. The uniform density has the lowest elasticity which numerically corresponds very closely to the elasticity of the representative-bank special case.

From Figure 20 we also see that the average markup is the highest for the Pareto economy, followed by the three other alternatives which are hard to distinguish from each other. The Pareto economy is by far the most concentrated of the four, and its largest banks choose abnormally

high credit markups. As a result, the aggregate markup gets very inflated. Finally, the average probability of default is the lowest in the Pareto economy, followed by uniform, baseline, and lognormal economies. The degree of concentration can be viewed as a good predictor of systemic stability: the Pareto economy is the most concentrated and is thus the least risky.

Overall, this exercise is a simple but useful demonstration of quantitative implications of the banking trilemma. The most concentrated economy, whose distribution is drawn from a Pareto density, is the most efficient, least competitive, and most stable. The prediction of our framework for the future of banking is thus the following. If banking concentration continues to go up, which seems to be a realistic assumption to make given the cost-cutting and competitive trends, then the macroeconomy will benefit from higher efficiency stemming from the right tail, will attain a greater buffer against financial crises, but will suffer from welfare losses due to rising financial markups. This prediction seems to be in line with the recent time-series experience of the U.S. banking sector: the industry has become more concentrated all the while markups have risen (Corbae and D'Erasmo, 2020a).

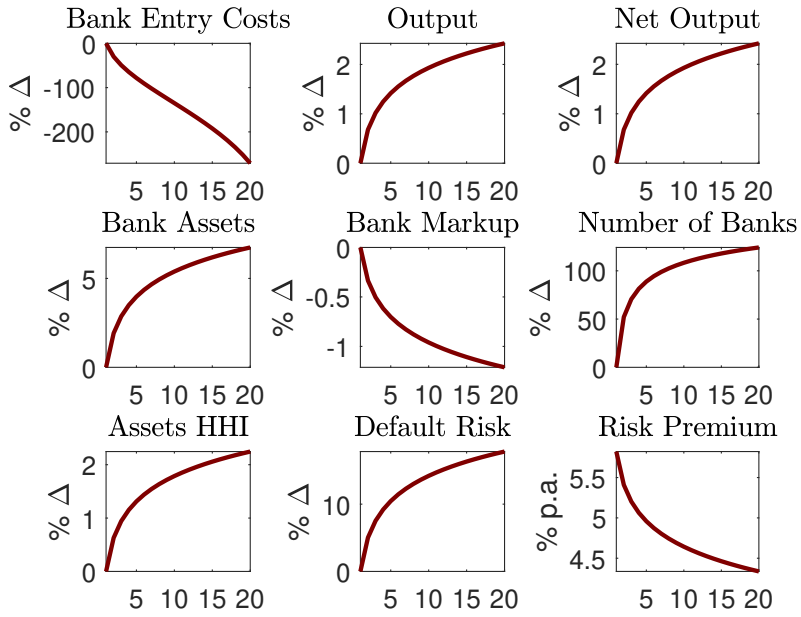
## A.2 Emergence of Fintech Credit

The global share of fintech in financial intermediary activities is growing rapidly, both in developed and developing financial markets (Claessens et al., 2018). In order to formalize the rise of fintech/bigtech firms, I extend the baseline model with endogenous bank entry in the spirit of Melitz (2003). There is now infinite mass of aspiring financiers who specialize in banking services. Before entry, every financier pays a fixed entry cost $e$ in units of capital. The rise of fintech will be simulated as a permanent decline in $e$. This is a reduced-form stand-in for various possible technological and preference-based explanations for this trend. Having paid the sunk cost, the financier receives an idiosyncratic return profitability draw $\xi_0 \in \Xi$ from the ergodic distribution $G_0(\xi)$ that is implied by the $\xi$ 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 $\{n_0, \xi_0\}$, the financier operates if and only if its expected discounted franchise value exceeds $e$. The value function of the entering financier is therefore:

$$V^e(n_0, \xi_0) \equiv \max \left[ V(n_0, \xi_0) - e, 0 \right] \tag{30}$$

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.

Figure 21: **Fintech Credit Growth**




Notes: Simulation of the fall in bank entry costs in the economy with endogenous entry.

The mass of financiers that decide to enter is $M$. The mass of active intermediaries, which now includes both incumbents and new entrants, is $H$. The stationary distribution of banks now keeps track of $M$ as well as the incumbents:

$$\mu'(\mathrm{n}',\xi') = \underbrace{\sum_{\xi} \mathrm{G}(\xi',\xi) \int \mathbb{1}_{\{(\mathrm{n},\xi)|\mathrm{K}(\mathrm{n},\xi)\in\mathbf{B}\}} \times \mathbb{1}_{\{\bar{\mathrm{d}}(\mathrm{n},\xi)=0\}} \eta(\mathrm{dn},\mathrm{d}\xi)}_{\text{Incumbents}} + \underbrace{\mathrm{M}' \int \mathbb{1}_{\{(\mathrm{n}_0,\xi)|\mathrm{K}(\mathrm{n},\xi)\in\mathbf{B}\}} \mathrm{G}_0(\xi)}_{\text{New Entrants}} \tag{31}$$

The law of motion of the distribution is now:

$$\eta_{\mathrm{t+1}}(\mathrm{n}_{\mathrm{t+1}},\xi_{\mathrm{t+1}}) = \Phi(\eta_{\mathrm{t}},\mathrm{M}_{\mathrm{t+1}}) \tag{32}$$

Credit market clearing now requires aggregate supply to equal demand from the incumbents and the financiers that wish to enter:

$$\underbrace{\mathrm{K}}_{\text{Aggregate Supply}} = \underbrace{\int_{\mathbf{B}} \Big(\mathrm{k}(\mathrm{n},\xi)\Big)\eta(\mathrm{dn},\mathrm{d}\xi)}_{\text{Incumbent Demand}} + \underbrace{\mathrm{M}\int_{\mathbf{B}} \Big(\mathrm{k}(\mathrm{n}_0,\xi_0)\Big)\mathrm{dG}(\xi_0)}_{\text{Entrants Demand}} + \underbrace{\mathrm{M}e}_{\text{Entry Cost}} \tag{33}$$

We set $e = 1.65$ for the baseline case. The fintech economy has $e = 0.11$. The number is calibrated such that the number of active banks in the economy roughly doubles.

Figure 21 shows the result of this exercise in the usual format. The model predicts that fintech credit will be responsible for more lending, output, and the number of active intermediaries - this is the direct extensive margin effect. Because the average intermediary is smaller, this lowers the average markup. We also observe a considerable elevation in financial fragility. Low entry costs essentially allow "too many" low-type lenders to enter every period by lowering the minimum profitability threshold below which financiers do not wish to stay. A growing mass of low-size, high-risk young intermediaries contributes to rising systemic fragility since default risk falls with net worth. This prediction is in line with the belief among regulators and policy-markers that fintech credit is a major source of financial stability for the 21st century.

## A.3 Intermediary Asset Pricing

Adrian et al. (2014) and He et al. (2016), among others, have popularized the intermediary asset pricing view: in contrast to conventional models, the true pricing kernel is a function of intermediary balance sheet ratios such as capital or leverage. This literature, however, relies predominantly on the representative agent assumption and abstracts from distributional dimensions.

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

$$\mathbb{E}_t\left[R^T_{t+1}(j) - R^{rf}_t(j)\right] = \underbrace{\frac{\lambda \overbrace{\varphi(j)}^{\text{Lagrange Multiplier}}}{\mathbb{E}_t\left(\hat{\Lambda}_{t+1}(j)\right)}}_{\text{Liquidity Premium}} + \underbrace{\nu(j)}_{\text{Default Premium}} \underbrace{+ \operatorname{cov}\left[-\frac{\hat{\Lambda}_{t+1}(j)}{\mathbb{E}_t\left(\hat{\Lambda}_{t+1}(j)\right)}, R^T_{t+1}(j)\right]}_{\text{Risk Premium}} \quad \forall j$$

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. If financial frictions are switched off, then the intermediation spread is zero. Excess returns in the baseline economy arise for two reasons. First, if the hard leverage constraint binds for any given bank j, or has a positive probability of binding in the future, then external funds are harder to obtain. This is the liquidity-induced external finance premium. Second, presence of bank default risk requires additional ex-ante compensation from the household's perspective. Note that the canonical risk premium is absent in the stationary equilibrium if we abstract from aggregate uncertainty.

Table 4 presents key asset pricing moments from the framework under different assumptions. Without any heterogeneity, the liquidity and default risk channels generate a premium of 2.28%.

Table 4: **Asset Pricing Moments**

| | Risk-Free Rate | Risky Return | Risk Premium |
|---|---|---|---|
| No Banks | 1.016 | 1.004 | 0 |
| Homogenous Bank | 1.016 | 1.038 | 0.023 |
| Only Monopolistic Competition | 1.016 | 1.037 | 0.021 |
| Only Idiosyncratic Risk | 1.025 | 1.060 | 0.035 |
| Baseline | 1.024 | 1.085 | 0.061 |

Notes: main asset pricing moments for various versions of the model. All percentages are annualized.

Adding monopolistic competition with variable markups and uninsurable idiosyncratic shocks gets us a large risk premium of 6.1%. This occurs because both liquidity and default risk premia are concentrated in the left tail of the distribution. Heterogeneity switches on the extensive margin, and a large equilibrium share of low-net-worth banks raises aggregate riskiness of the economy. In addition, with monopolistic competition relative prices fall heavily with bank net worth - smaller banks are less competitive in price terms. Overall, without aggregate uncertainty and relying solely on idiosyncratic shocks and the structure of the model we therefore can explain essentially all of the unconditional risk premia observed in U.S. data.

# B Targeted Stabilization Policies

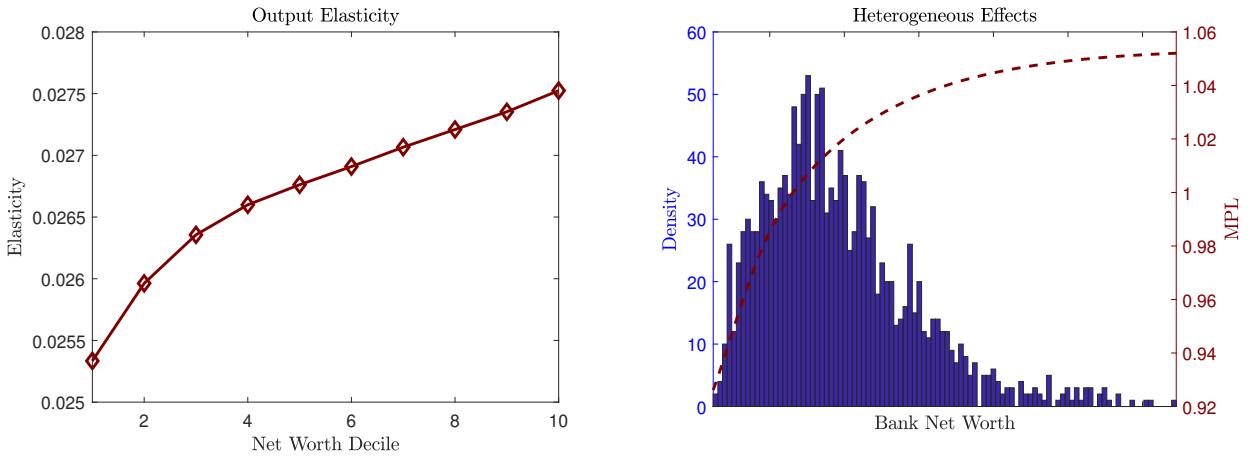
We now show how in our framework one can easily analyze targeted or bank-level regulatory interventions. We focus on equity injections and liquidity facilities.[2]

## B.1 Equity Injections

Credit policy has been modelled in several representative-agent Macroeconomic frameworks, for example He and Krishnamurthy (2013) and Curdia and Woodford (2016). We move beyond aggregate credit policy analysis and estimate conditional macro elasticities when equity injections are allowed on any *individual* bank in the distribution. We proceed in two steps. First, we break the distribution of bank net worth into ten bins (deciles). For each decile $\iota = 1 \ldots 10$, we assume that the government increases by one percent the net worth of each bank in $\iota$ but not anywhere else in the

[2]I also consider two additional policy types. First, targeted lending facility. This is a scenario when the monetary authority takes over market lending on behalf of the intermediaries. In the model, this corresponds to the market for differentiated capital goods. This policy alters the distribution of marginal costs in the banking sector such that the cost of funds of the central bank is lower than of any bank in the ergodic distribution. Second, targeted bank-level guarantees. This exercise supplements the deposit insurance scheme from Section 5.2 which was an aggregate policy. Results for targeted direct lending and bank-level debt guarantees are available upon request.

Figure 22: **Macroeconomic Effects of Targeted Equity Injections**




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

economy. Second, we compute the macro elasticity with respect to targeted policies by integrating 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 separate times.

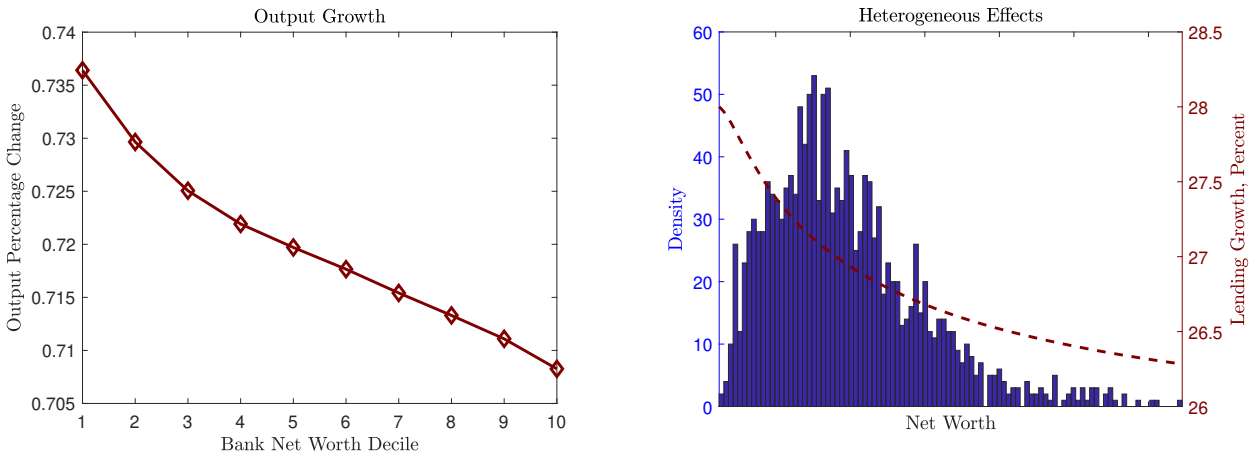
Figure 22 plots the result. We observe that there are efficiency gains from injecting equity into large intermediaries. The elasticity of aggregate output with respect to decile-specific credit policies is an upward-sloping line. This result is driven by the shape of the MPL distribution - larger banks 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.[3]

## B.2 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 - $\lambda$ - 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 $\lambda$ on behalf of private lenders. In particular, we allow $\lambda$ to be relaxed on *any* bank in the distribution. In practice, this intervention can be mapped to

[3]These bailouts are unexpected and do not generate additional moral hazard frictions ex-ante. The implicit bailout subsidy is internalized in an earlier Section 5.5.

Figure 23: **Macroeconomic Effects of Targeted Liquidity Facilities**




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

discount window lending to banks secured by the credit portfolio.

In order to facilitate the cleanest possible analysis, we assume that the leverage constraint binds on all banks in the distribution.[4] With the binding leverage constraint, it is straightforward to solve for the bank-specific leverage ratio:

$$\phi(j) = \frac{\nu_a(j)}{\lambda - \mu_a(j)} \tag{34}$$

where, as before, $\phi(j)$ is market leverage, $\nu_a(j)$ is the discounted cost of bank liabilities, $\mu_a(j)$ are excess returns on the risky asset. Notice how according to this formula, relaxation of liquidity conditions (as proxied by a reduction in $\lambda$) 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 $\lambda_\iota$ on decile $\iota = 1 \ldots 10$ of the banks net worth distribution by 10% relative to the baseline value of 0.1. The exogenous shock is thus invariant to the region of the distribution which is targeted. The only variable parameter in this policy intervention is the decile of the bank net worth distribution. For each of the ten policy counterfactuals, we compute the conditional output elasticity. Figure 23 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 $\lambda_\iota$ on $\phi_\iota$ is negative and declining with bank size due to diminishing marginal costs of funds.

[4]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.

# C MIT Shocks to Aggregate Productivity

In this section we study the transmission mechanism of exogenous, unanticipated aggregate shocks to Total Factor Productivity ($A_t$). After a sudden one standard-deviation decline, $A_t$ reverts back to the steady state with an autoregressive factor of 0.6. We assume that any policy interventions are fully unanticipated and occur only during crisis episodes and never in the steady state or when productivity is high.

We are interested in tracking the responses of all aggregate quantities and prices but focus on aggregate demand $K_t$ for compactness. Let us write $K_t$ as an explicit function of the exogenous transitory shock, equilibrium prices, and policy interventions $\{\Omega_t\}_{t\geq 0}$, with $\{\Omega_t\} = \{R_t^k, \bar{R}_t, P_t, \tau_t\}$ and where $\tau_t$ summarizes any policy actions of the government:

$$K_t\Big(\{\Omega_t\}_{t\geq 0}\Big) = \int k_t\Big(n, \xi; A_t, \{\Omega_t\}_{t\geq 0}\Big)\mu_t(dn, d\xi) \tag{35}$$

where $k_t\Big(n, \xi; A_t, \{\Omega_t\}_{t\geq 0}\Big)$ is the bank-level policy function for bank credit (assets). Recall that $\mu(n, \xi)$ is the joint distribution of bank net worth and idiosyncratic rate of return risk.

We can decompose the total response of credit supply at $t = 0$ by differentiating Equation 35:[5]

$$dK_0 = \Big[\underbrace{\int_0^\infty \frac{\partial K_0}{\partial A_t} dA_t}_{\text{Direct Effect}} + \underbrace{\int_0^\infty \Big(\frac{\partial K_t}{\partial \bar{R}_t} d\bar{R}_t + \frac{\partial K_t}{\partial R_t^k} dR_t^k + \frac{\partial K_t}{\partial P_t} dP_t + \frac{\partial K_t}{\partial \tau_t} d\tau_t\Big)}_{\text{Indirect Effect}}\Big] dt \tag{36}$$

The first term in Equation 36 summarizes direct effects of the shock to productivity on credit supply, while holding all aggregate prices and policies constant. All banks in the distribution respond to $A_t$ directly because aggregate productivity impacts the path of aggregate returns on investment, which enters explicitly the law of motion of bank net worth through $R_t^T(j)$. The direct effect can be further decomposed into the cross section of bank-level marginal propensities to lend:

$$\int_0^\infty \frac{\partial K_0}{\partial A_t} dA_t dt = \int_0^\infty \Big[\int \frac{\partial k_0(n, \xi; A_t, \{\bar{\Omega}_t\}_{t\geq 0})}{\partial A_t} \bar{\mu}(dn, d\xi)\Big] dA_t dt \tag{37}$$

With $\bar{\Omega}$ and $\bar{\mu}$ fixed at the steady-state values. That is, the total direct effect comprises the aggregated partial-equilibrium response of credit supply to the exogenous disturbance alone without updating aggregate general equilibrium variables and the banking distribution.

The indirect effect from Equation 36 includes four distinct channels of transmission. First, aggregate productivity impacts demand for investment. Because firms require external financing

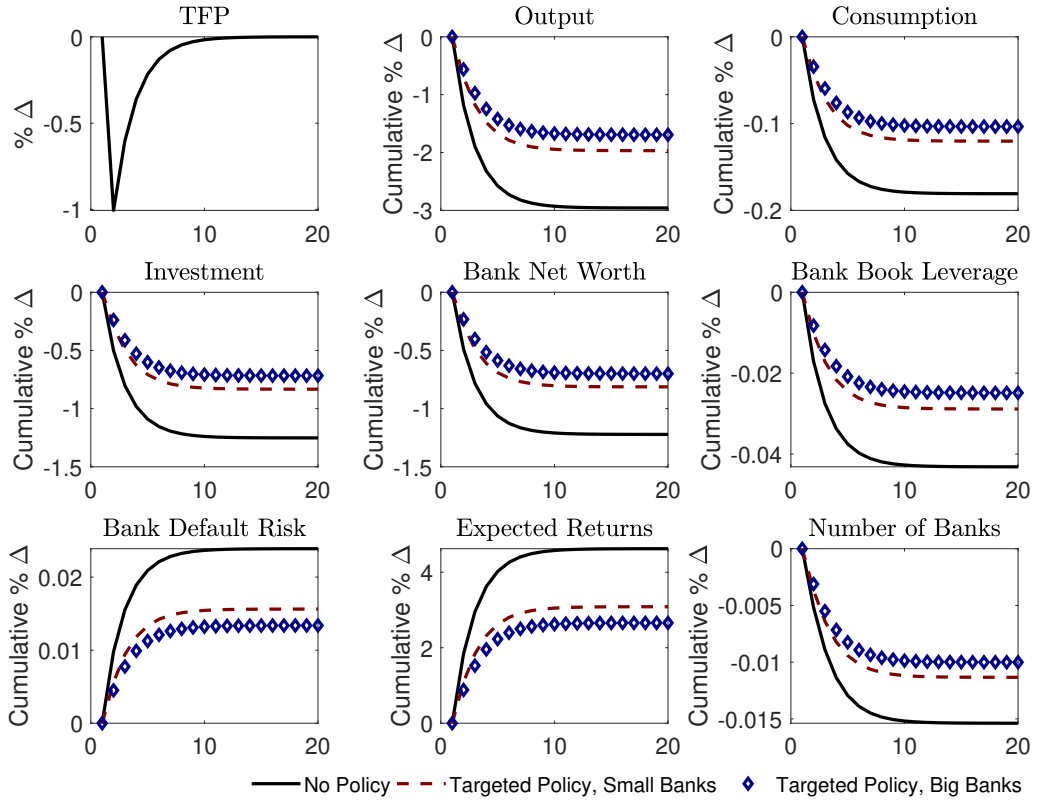
[5] Our decomposition is similar to the one applied in Kaplan et al. (2018) who study distributional implications in the response of aggregate consumption to monetary policy shocks.

in order to produce, this immediately translates into the demand for bank lending activities. Banks, because of credit market power, respond to increased demand by adjusting their private, bank-level markups and prices. In addition, prices adjust also because the distribution of bank net worth shifts and, as we concluded in previous sections, relative prices and marginal costs vary with net worth. In the aggregate, this moves $P_t$, which further feeds into bank-level choices of credit supply. Recall that banks do not internalize this GE channel, which is an aggregate credit supply externality.

Second, banks react to movements in investment demand by requesting more or less short-term debt from the households. In the deposit market, this affects the distribution of deposit interest rates, which drives the aggregate rate $\bar{R}_t$. Third, every second-level general equilibrium channel feeds into the aggregate stock of capital which, together with the aggregate price, determines the new level of systematic returns $R_t^k$.

Finally, banks will react to policy interventions from the fiscal and monetary authorities, if there is any. In previous sections, we discussed systematic and targeted (bank-level) equity injections and liquidity facilities. All these policies are summarized in the term $\tau_t$, which is understood to be capturing any aggregate or bank-level policy responses. Credit policies of any kind will perturb allocations in the banking sector one way or another. Equity injections induce direct credit supply responses because those explicitly augment one of the idiosyncratic states of the banking problem - $n_t(j)$. Liquidity facilities impact the probability of the leverage constraint binding in the future, which weighs in on the banks' decision to take on more or less balance sheet risk.

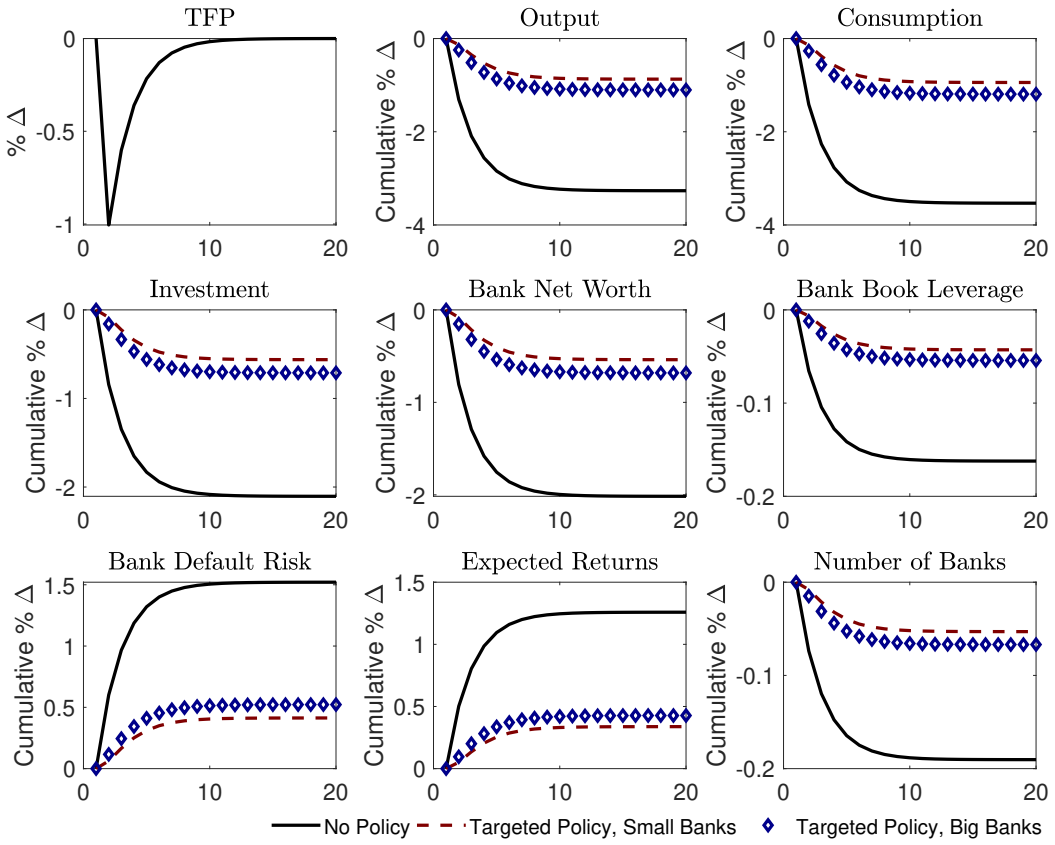
Figure 24: **Crisis Experiment: Targeted Equity Injections**




Notes: Responses to a one standard-deviation negative shock to $A_t$, with and without targeted equity injections. Baseline economy.

We begin the presentation of numerical results with our baseline economy that goes through an aggregate economic crisis but does not get a response from fiscal/monetary institutions. Figure 24 portrays the results. For all variables except the transitory $A_t$ shock, we present cumulative impulse response functions. We observe that the economy is going through a contraction of aggregate consumption, output, and investment of the magnitudes that are similar to the 2007-2009 Great Recession. Bank net worth and book leverage fall. Bank balance sheets become more risky as the aggregate (average over the entire distribution) probability of insolvency risk increases. As the average bank in the distribution is smaller in terms of net worth, the leverage constraint binds or is more likely to bind for a larger fraction of the intermediaries. This translates into the rise of equilibrium excess returns. Finally, because bank franchises decline in value, fewer banks decide to enter and the number of active intermediaries falls.

Figure 25: **Crisis Experiment: Liquidity Facility**




Notes: Responses to a one standard-deviation negative shock to $A_t$, with and without targeted liquidity facilities.

Figure 24 also plots impulses and responses under targeted, bank-level policy interventions. First, we look at direct equity injections into only small or only large banks, and compare the response functions. We assume that the government increases net worth of every bank in the targeted mass by 10%. We define "small" and "large" banks as those intermediaries whose net worth is in the bottom and top deciles of the steady-state, ergodic distribution of bank net worth, respectively. We see that equity that is injected into big banks has a bigger bang for the buck than the equivalent investment into small banks. This is due to the positive slope of the MPL schedule and a bigger macro elasticity. Large banks have a greater propensity to lend than small banks because of lower marginal costs and economies of scale..

Figure 25 plots the same numerical experiment but now with targeted liquidity facilities. These policies reduce the fraction of divertible assets $\lambda$ by 10% for a certain decile of the bank net worth distribution. We see that discount-window-based lending considerably dampens recessions, particularly if applied to small banks. This is due to the leverage constraint binding much more

often for banks with low levels of net worth, particularly in recessions when net worth is low. Any policy that reduces $\lambda$ induces a greater credit supply response if directed to the agents that are affected by the moral hazard friction by more. Because now the risk of a tightening leverage constraint relatively dissipates, excess returns increase by less, which leads to lower risks of default, more lending, and a relatively stronger macroeconomic responses (the economy is still contracting but the cumulative magnitude is lower).

# D Additional Model Details and Derivations

## D.1 Bank Scale Variance

In this section we demonstrate how the baseline economy features scale variance and nests the representative-bank special case. We visualize the mechanism graphically on figure 26. We analyze the optimal choice of bank market leverage $\frac{pk}{n}$ in three different situations. First, we start with the representative-bank case with complete markets ($\sigma_\xi$=0 and $\kappa$=0), and linear non-interest expenses ($\zeta_2 = 1$). As can be seen from the figure, linearity and complete markets make the leverage ratio one-dimensional and independent of the state of initial net worth. Second, the downward-sloping line on the left panel of Figure 26 plots optimal leverage for an extension that allows for scale variance ($\zeta_2 > 1$). Notice how leverage is now decreasing in net worth. Finally, in the right panel of the Figure, we relax the assumption of market completeness. Moreover, because we continue to retain scale variance, the optimal leverage ratio now depends on two states: $\xi(j)$ and n(j): low-n(j), high-$\xi$(j) banks choose the highest leverage in the economy.

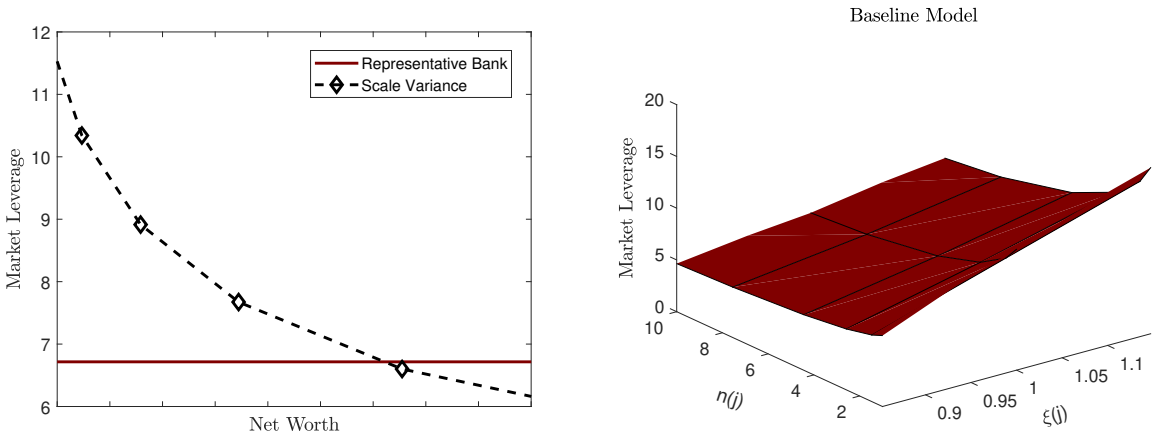
An important feature of this class of models with financial intermediaries is linearity with respect to net worth. This assumption normally allows the model to be aggregated explicitly. I can formalize the departure from homogeneity by formally proving that the value function of the bank in our model 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), among many others. Proposition 2 formalizes the intuition.

**Proposition 2** (Bank Scale Variance)**.** *The solution to the incumbent banker's problem, conditional on initial net worth* n(j) *and idiosyncratic return* $\xi$(j)*, is*

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

*where the marginal value of net worth is:*

Figure 26: **Bank Scale Variance**




Notes: Left picture shows how bank leverage depends on net worth in two regimes. "Representative bank" is the case without idiosyncratic shocks and scale variance ($\zeta_2 = 1$). "Scale variance" is the case with scale variance ($\zeta_2 > 1$) and no idiosyncratic shocks. Right picture shows how bank leverage depends on net worth in the baseline economy with both scale variance and idiosyncratic shocks.

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

*and the multiplier on the moral hazard leverage constraint is*

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

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

$$\mathbb{E}\left[\tilde{\Lambda}\left(\mathrm{R}^{\mathrm{T}}(\mathrm{j})-\bar{\mathrm{R}}(\mathrm{j})\right)\right] = \lambda\varphi\Big(\mathrm{n(j)},\xi(\mathrm{j})\Big)$$

$$\varphi\Big(\mathrm{n(j)},\xi(\mathrm{j})\Big)\Big[\vartheta\Big(\mathrm{n(j)},\xi(\mathrm{j})\Big)-\lambda\phi(\mathrm{j})\Big] = 0$$

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

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

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

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

And the Lagrange multiplier on the leverage constraint is

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

Note that when $\epsilon$ = then market leverage becomes $\phi(j)=k(j)^{\frac{\theta-1}{\theta}}K^{\frac{1}{\theta}}Pn(j)^{-1}$. The guess is verified if $\varphi\Big(n(j),\xi(j)\Big)<1$. Net worth-dependency is guaranteed by $\zeta_2\neq 1$ (for a given $\zeta_1\neq 0$) so that each bank with a different n(j) chooses its own leverage ratio $\phi(j)$. Furthermore, with $\kappa>0$, $\phi(j)$ also explicitly depends on $\xi(j)$. As a result, explicit aggregation in the banking sector is not possible as the linearity condition is not satisfied. Financial intermediaries are ex-post heterogeneous in terms of returns, which feeds into all other balance sheet and income statement characteristics because of scale dependency.[6] □

## D.2 Bank Markups and Marginal Costs

In this section we provide a proof for Proposition 1. 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 $\zeta_1=\zeta_2$.

Show that the bank price-setting rule is:

$$\frac{p(j)}{P}=\mu(x)\frac{k(j)^{\zeta_2-1}}{R^T(j)-\bar{R}(j)}$$

Each bank j solves

$$\max_{k(j)}\left\{\tilde{\Lambda}\Big(1-\upsilon(j)\Big)\Big[R^T(j)p(j)k(j)-\bar{R}\Big(p(j)k(j)-n(j)\Big)-\frac{1}{\zeta_2}k(j)^{\zeta_2}\Big]\right\}\quad\text{s.t.}\quad p_t(j)=\Upsilon'\left(\frac{k(j)}{K}\right)Z_t$$

[6] Note that $\{\epsilon,\theta\}$ do not impact scale-dependency but do change the level and curvature of the $\vartheta\Big(n(j),\xi(j)\Big)$ surface.

where $Z \coloneqq \left(\int_0^1 \Upsilon'\left(\frac{k(j)}{K}\right)\frac{k(j)}{K}dj\right)^{-1}$. The first order condition is

$$\tilde{\Lambda}\Big(1-\nu(j)\Big)\left\{\Big(R^T(j)-\bar{R}(j)\Big)\Big(p(j)+k(j)\frac{\partial p(j)}{\partial k(j)}-k(j)^{\zeta_2-1}\Big)\right\}=0$$

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

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

where x is relative bank size. The markup function $\mu(x)$ is:

$$\mu(x)=\frac{\theta x^{-\frac{\epsilon}{\theta}}}{\theta x^{-\frac{\epsilon}{\theta}}-1}$$

The marginal cost MC(j) is given by:

$$MC(j) \coloneqq \frac{1}{R^T(j)-\bar{R}(j)}k(j)^{\zeta_2-1}$$

The price-setting rule given marginal costs is thus:

$$\frac{p(j)}{P}=\frac{\theta x^{-\frac{\epsilon}{\theta}}}{\theta x^{-\frac{\epsilon}{\theta}}-1}\frac{k(j)^{\zeta_2-1}}{R^T(j)-\bar{R}(j)}$$

where the first term on the right hand side is the markup and the second term is the marginal cost.

**Constant Markup** Whenever $\epsilon = 0$ the relative price price rule becomes:

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

where $\frac{\theta}{\theta-1}$ is the constant markup over the marginal cost which is now:

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

Solving out aggregate prices gives:

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

Note that this equation resembles the canonical price rule in Blanchard and Kiyotaki (1987).

□

# E Microfoundations and Extensions

## E.1 Discrete Choice Microfoundation

This section provides a brief theoretical foundation for the representative-agent capital goods producer's monopolistically competitive credit demand system. My approach follows closely Anderson et al. (1989). We focus on the analytically more convenient case when $\epsilon = 0$. 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 $\epsilon_{ij}$ 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.

$$Q_i(\bar{\epsilon}; k_i) = \underbrace{\log k_i + Y - p_i k_i}_{\text{Homogenous across j}} - \underbrace{\sum_{k=1}^{M} \left(\bar{\epsilon}^k - \epsilon_i^k\right)^2}_{\text{Heterogeneous across j}} \qquad i=1 \ldots H \tag{38}$$

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:

$$\bar{E}^j = \frac{\log \frac{p(j)}{p_n}}{4} \tag{39}$$

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:

$$k_n = \frac{1}{p_n} \int_{-\infty}^{\bar{\epsilon}_1} \ldots \int_{-\infty}^{\bar{\epsilon}_{n-1}} f\left(\bar{E}^j\right) d\bar{\epsilon} \tag{40}$$

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. $p_1, \ldots, p_{H-1}$:

$$\frac{\partial^{H-1} k_n}{\partial p_1 \ldots \partial p_{H-1}} = \frac{1}{p_1 \ldots p_n} 4^{1-H} f\left(\bar{E}^j\right) \tag{41}$$

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:

$$f(\bar{\epsilon}) = H\frac{4}{\bar{\theta}}^{H-1} (H-1)! \frac{\prod_{i=1}^{H-1} \exp\left(-4/\bar{\theta}\epsilon^i\right)}{\left[1 + \sum(j)^{H-1} \exp\left(-4/\bar{\theta}\epsilon^j\right)\right]^H} \tag{42}$$

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

$$\frac{\partial^{H-1} k_n}{\partial p_1 \ldots \partial p_{H-1}} = H\bar{\theta}^{1-H} (H-1)! \frac{\prod(j)^H p(j)^{-1/\bar{\theta}-1}}{\left[\sum(j)^H p(j)^{-1/\bar{\theta}}\right]^H} \tag{43}$$

Integrating gives us optimal credit demand

$$k_n = H p_n^{-1/\bar{\theta}-1} \left[\sum(j)^H p(j)^{-1/\bar{\theta}}\right]^{-1} \tag{44}$$

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:

$$k(j) = \left(\frac{p(j)}{P}\right)^{-\theta} K \tag{45}$$

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 $\theta$. 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 $\bar{\theta} \to \infty$ we recover a case with a single pure monopoly provider of credit. As $\bar{\theta} \to 0$ 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.

## E.2 Portfolio Returns

In this section we explain how our formulation of total portfolio returns (Equation 8) can be microfounded. Suppose there are N banks and credit markets. These credit markets could be understood in at least three different ways: units in geographical space (counties), segmented industries, or segmented financial varieties (products/services). The model is isomorphic to any of these interpretations. Now suppose that each bank $b$ specializes in one credit market $c$ and overweighs it by $0 < \kappa < 1$. Concentration can be motivated by a variety of theories, including but not limited to "home" bias in bank lending (Juelsrud and Wold, 2020) or asymmetric information (Van Nieuwerburgh and Veldkamp, 2009). Assume that market-specific returns $R^j$ are not diversifiable/insurable. This assumption can be motivated by the empirical findings in Galaasen et al. (2020). Then, the bank-specific portfolio return can be written as:

$$R^b = \sum_j^N \frac{1}{N} R^j + \kappa R^c - \kappa R^{-c} \tag{46}$$

where $R^{-c}$ is the return on a portfolio that excludes the bank's favorite market c. Now, we assume that N is large enough such that $R^{-c}$ is approximately equal to the return on the market portfolio $R^k$. That is, credit markets are atomistic:

$$R^b \approx R^k + \kappa R^c - \kappa R^k = \kappa R^c + (1-\kappa) R^k$$

Which is the same formulation that we used in Equation 8, except that in the model $R^c$ is $\xi(j)$ and follows an autoregressive process. Now, total return across all banks can be written as:

$$R^{\text{total}} = \sum_b^N \frac{1}{N} \kappa R^c + (1-\kappa) \sum_b^N R^k = R^k$$

That is, in the aggregate, credit market-specific idiosyncratic returns vanish and banks are exposed only to the systematic component of returns $R^k$. What makes idiosyncratic return risk an intertemporal problem for banks are (a) scale variance and (b) persistence of $\xi(j)$.

## E.3 Two-Sector Extension

The baseline economy in the main text features a single capital goods sector which is intermediated by imperfectly competitive banks. It's possible to generalize our setup to two types of capital goods. Suppose the first capital good $K_{at}$ is imperfectly differentiated across the mass of banks $H_t$. These are the financial varieties which we discuss in main text. The second good type $K_{bt}$ is a perfect substitute across lenders. This proxies standard fixed-term commercial loans which

are homogenized across banks, who in turn face perfect competition in this market. We continue to assume that there is a representative capital goods producer that is financially constrained and requires bank funds in order to produce the capital stock. The production stage of the capital stock now consists of two steps. First, we determine the equilibrium fraction of differentiated capital goods $K_{at}$. The capital goods firm solves the following problem:

$$\min_{K_{at},K_{bt}} P_t K_{at} + K_{bt} \quad \text{s.t.} \quad K_{at}^{\chi} K_{bt}^{1-\chi} = K_t \tag{47}$$

Where $0 < \chi < 1$ is the elasticity of substitution across the types of capital goods. The solution delivers a set of two familiar equations: $P_t K_{at} = \chi K_t$ and $K_{bt} = (1-\chi)K_t$. That is, the share of financial varieties in the economy is time-invariant and is equal to $\chi$. The second stage of the problem is determination of the demand for individual varieties $k_t(j)$ within the $K_{at}$ sector.

The parameter $\chi$ could be taken to the data and mapped to the scale and intensity of shadow banking activities before the Crisis (Gorton and Metrick, 2010). Parameter statics in $\chi$ could be used to simulate advancements in financial innovation and/or the rise of complexity in the credit market.

# F Numerical Solution Algorithm

In this section we lay out the numerical algorithm that is used to solve different variants of the model. We first describe how to solve the baseline unregulated market economy. We then show how to solve for constrained efficient allocations of the social planner and how to decentralize them.

## F.1 Unregulated Market Equilibrium

Below we list 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 - Stationary Industry Equilibrium**

1. Guess some initial values for aggregate endogenous states $\{K, P, \Lambda\}$. Compute $R^k$. Guess some initial values for idiosyncratic endogenous states $\{\upsilon(j), \bar{R}(j)\}$
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=2,000 periods.

4. Solve the new entry problem, if entry is endogenous. Determine the mass of entrants and their aggregate demand for capital in each period of the simulation.

5. Compute economywide new guesses for aggregate $K'$ and $P'$. Construct a new $R^{k'}$. Check if $K'$ is sufficiently close to K. If not, return to Step 2. If yes, continue with the program.

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

7. Solve the household's problem using time iteration. Get new $\Lambda'$.

8. Compare $\{\bar{R}(j)\}$ with $\{\bar{R}'(j)\}$, K with K', and P with P'. If maximal errors are within the tolerance level, general equilibrium is solved. If not, update $\{\bar{R}(j)\}$, K, and P. Return to Step 2 and continue the iteration.

We require convergence tolerances of $10^{-6}$ for general equilibrium deposit rates, $10^{-5}$ for the bankers' and household's problems, and $10^{-3}$ for aggregate capital and prices.[7]

## F.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), $\xi(j)$, and the choice of k(j). That is:

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

Once the optimal k(j) is found, that gives us p(j) through the credit demand function and d(j) from the balance sheet constraint. 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.

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

**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 $\{\upsilon(j), \bar{R}(j)\}$. The 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=2,000 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 p(j). Compute new guesses for $\{\alpha^{k'}, \alpha^{p'}\}$.

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

6. Compute economywide new guesses for $K'$ and $P'$. Construct a new $R^{k'}$. If K' and P' are sufficiently close to K and P, respectively, then continue. If not, return to Step 2.

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 $\Lambda'$.

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, the constrained efficient equilibrium is solved. If not, update $\{\alpha^{k}, \alpha^{p}\}$ and $\{\bar{R}(j)\}$. Return to Step 3 and continue the iteration.

We decentralize constrained efficient equilibria with size-dependent taxes on bank gross returns $\tau\Big(n(j), \xi(j)\Big)$. In each iteration, we solve the financial intermediary problem subject to a conjectured

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 at their constrained-efficient levels.

# G Data Description

Empirical data used for model validation is obtained from the U.S. Call Reports. Table 5 details the definition of every variable used. Our quarterly sample is 2010q1-2019q4. All variables are truncated at the 1% and 99% levels. Model variables are defined as stated in the Table and obtained from a stochastic simulation of the stationary industry equilibrium with N=1 intermediaries and T=2,000 quarters.

Table 5: **Description of Financial Variables**

| Data | | |
|---|---|---|
| Variable Name | Description | Source |
| Assets | Total assets (RCFD2170) | Call reports |
| Equity | Total assets (RCFD2170) - total liabilities (RCFD2948) | Call reports |
| Leverage ratio | Assets / equity | Call reports |
| Deposit expenses | Interest expense on domestic deposits. Equals total interest expense on deposits (RIAD4170) - interest expense on foreign deposits (RIAD4172) | Call reports |
| Non-interest expenses | Total noninterest expenses (RIAD4093) | Call reports |
| Net interest income | Net interest income (RIAD4074) | Call reports |
| **Model** | | |
| Variable Name | Description | |
| Assets | $k(j)$ | |
| Equity | $n(j)$ | |
| Leverage Ratio | $\frac{k(j)}{n(j)}$ | |
| Deposit expenses | $\bar{R}(j)d(j)$ | |
| Non-interest expenses | $\frac{1}{\zeta_1}k(j)^{\zeta_2}$ | |
| Net interest income | $R^T(j) - \bar{R}(j)$ | |

Notes: This table details the construction, definition, and sourcing of all empirical and model variables used for parameterization and validation.

# References


Adrian, T., E. Etula, and T. Muir (2014): "Financial Intermediaries and the Cross-Section of Asset Returns," *Journal of Finance*, 69(6), 2557–2596.

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).

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

Claessens, S., J. Frost, G. Turner, and F. Zhu (2018): "Fintech credit markets around the world: size, drivers and policy issues," *BIS Quarterly Review*.

Constancio, V. (2016): "Challenges for the European Banking Industry," *Conference on "European Banking Industry: what's next?"*.

Corbae, D. and P. D'Erasmo (2020a): "Capital Requirements in a Quantitative Model of Banking Industry Dynamics," *NBER Working Paper*, 25424.

——— (2020b): "Rising bank concentration," *Journal of Economic Dynamics and Control*, 115.

Curdia, V. and M. Woodford (2016): "Credit Frictions and Optimal Monetary Policy," *Journal of Monetary Economics*, 84.

Gabaix, X. (2009): "Power Laws in Economics and Finance," *Annual Review of Economics*, 1.

Galaasen, S., R. Jamilov, R. Juelsrud, and H. Rey (2020): "Granular Credit Risk," *NBER Working Paper 27994*.

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

Gorton, G. and A. Metrick (2010): "Regulating the Shadow Banking System," *Brookings Papers on Economic Activity*, Fall.

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

He, Z. and A. Krishnamurthy (2013): "Intermediary Asset Pricing," *Journal of Financial Economics*, 103(2), 732–770.

Juelsrud, R. E. and E. G. Wold (2020): "Risk-weighted capital requirements and portfolio rebalancing," *Journal of Financial Intermediation*, 41, 100806.

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

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

Van Nieuwerburgh, S. and L. Veldkamp (2009): "Information immobility and the home bias puzzle," *The Journal of Finance*, 64, 1187–1215.