London Business School

Essays in Financial Macroeconomics

Rustam Jamilov

A thesis submitted to the London Business School for the Degree of Doctor of Philosophy

June 2021

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London Business School is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without the prior written consent of the author. I warrant that this authorization does not, to the best of my belief, infringe the rights of any third party.

Statement of conjoint work

Chapter 2 is joint work with Sigurd Galaasen, Ragnar Juelsrud, and Hélène Rey. I contributed 25% of this work.

Chapter 3 is joint work with Tommaso Monacelli. I contributed 50% of this work.

Abstract

This dissertation explores the relationship between financial frictions and the real economy. It studies three channels of macro-financial transmission: market power of financial intermediaries, credit portfolio risk that originates from granular borrowers, and aggregate implications of countercyclical bank income risk.

The first Chapter of the dissertation develops a quantitative model with financial intermediaries that are heterogeneous along four empirically-motivated dimensions: balance sheet size, credit market power, default risk, and efficiency. The framework highlights a *trilemma* for bank regulation: financial competition, efficiency, and stability are incompatible – no policy intervention can improve all three facets simultaneously. This trilateral trade-off is shown to extend to numerous classic and new issues in macroeconomics and banking such as deposit insurance, capital requirements, the "too-big-to-fail" hazard, emergence of fintech lending, etc.

The second Chapter (joint with Hélène Rey, Ragnar Juelsrud, and Sigurd Galaasen) provides the first bottom-up quantification of *single-name* credit concentration risk by applying a novel empirical approach to administrative matched bank-firm data from Norway. The study exploits granularity in the distribution of loan shares and shows that idiosyncratic shocks to large borrowers survive portfolio aggregation and impact bank outcomes, contrary to conventional theories. Moreover, granular credit shocks spill over to other firms through banks' balance sheets, increase probability of default for the affected firms, thus having a significant impact on the real economy.

The third Chapter (joint with Tommaso Monacelli) builds a quantitative macroeconomic model with aggregate uncertainty and heterogeneous financial intermediaries that face *counter-cyclical* idiosyncratic rate of return shocks. Counter-cyclicality of bank income risk is estimated directly from U.S. bank-level data, holds for the past 40 years, and extends to the most recent COVID-19 recession. The channel is found to significantly amplify and prolong economic downturns.

Acknowledgments

I am indebted to many people, without whom my Ph.D. journey would not have been possible.

First and foremost, I would like to thank Hélène Rey, my advisor. Meeting Hélène has been the greatest honor and privilege of my career. Hélène is an extraordinarily thoughtful academic with an incredible economic intuition, second-to-none attention to detail, an astounding breadth of knowledge, and highest levels of professional courtesy, ethics, and integrity. The span of her expertise is remarkable and something I have been privileged to benefit from as an aspiring scholar. Be it international finance, general equilibrium macroeconomics, monetary policy, or online machine learning: every meeting and conversation with her is a unique learning experience. She is also the most hard working person I have ever met. Either working on the next great paper, advising global leaders on financial stability or climate change risk, or teaching a new course, Hélène's office lights are always on. Hélène's devotion to her students is unparalleled. She has always been very supportive, empowering, and generous with her time. I feel so grateful that she accepted me as one of her advisees and am very much looking forward to our future interactions.

I have benefited tremendously from many useful conversations with faculty at LBS. Elias Papaioannou has been very instrumental for my career as a teacher and advisor. I thank Vania Stavrakeva for numerous comments and suggestions that have greatly improved my job market paper. Richard Portes has been extremely kind with his encouragement and uplifting feedback. Francisco Gomez, Joseba Martinez, Andrew Scott, Paolo Surico have been a constant source of support and general life advice which have helped me become who I am as a researcher. Over the years I have also benefited tremendously from numerous interactions with Jean-Pierre Benoît, Andrea Galeotti, David Myatt, Emre Ozdenoren, Lucrezia Reichlin, and Linda Yueh. Last but not least, special thanks to Judith Fry, Sue Haines, and Mariana Uchida for running the best programme in the world. Huge thanks to Eleanor Burke for being the backbone of the economics department.

I am very fortunate to work with several fantastic co-authors on various projects. Tommaso Monacelli was my undergraduate teacher and thesis advisor at Bocconi. The first thesis that I wrote under his guidance was an analytical and qualitative extension of the IS-LM framework to the zero-lower bound environment. Unfortunately, I lost the original .doc file, otherwise it could have become a nice paper. I was told to write a second one, I think because the draft had zero equations. So I wrote a second dissertation from scratch and in two weeks. The submission did get a grade of 30 out of 30 in the end. Fast forward almost 10 years and remarkably enough we are working together on "Bewley Banks" and other projects, to which I am very much looking forward. I want to thank Sigurd Galaasen and Ragnar Juelsrud for taking a chance on me, agreeing to work together, and being the best hosts in Oslo. I have learnt a great deal from them and am tremendously excited about our future collaborations. A huge thanks to Ahmed Tahoun for playing an instrumental part in the creation and execution of our "The Anatomy of Cyber Risk" paper. I have learnt so much from him and benefited from his always kind, patient, and very thoughtful ideas and feedback.

I have had a fantastic group of classmates at LBS. I have benefited from continuously great conversations about economics with Luis Fonseca, Michele Andreolli, Diego Kaenzig, Tsvetelina Nenova, Ashley Perry, Juan-Antolin Diaz, Natalie Rickard, Jinglun Yao, Johan Moen, Guy Pincus, and many others. Outside of the economics department I have also had the pleasure of meeting and frequently talking with Shikhar Singla, Iris Steenkamp, Thilo Kind, Nazli Sonmez, Aldona Kapacinskaite, Nazli Gurdamar, Esben Christensen, Trevor Young, Varun Sharma, and many more.

I extend my gratitude to Kasper Roszbach and everyone at Norges Bank for amazing hospitality. Special thanks to the Wheeler Institute for Business and Development for support, and especially to Trias Kolokitha-Schmitz and Tiago Ivo Martinho. Huge thanks to the AQR Asset Management Institute for support, and especially to Anna Pavlova, Stephen Schaefer, and Eleanor Eiserman.

I owe everything to my family. My wife, Jamila, is my pride and an eternal source of inspiration. Thank you for all your sacrifices and everything you have done: relocating to the UK with me, leaving family and friends behind, mastering a foreign language, raising our daughter, finishing your degree, finding a job, then finding another job, and then finding an even better job, and then supporting me financially for most of the latter months of my PhD all the while managing to feed and keep me and the child alive. I would not have achieved anything in life without my parents, Mahammad and Nigar, my brother Zaur, and my beloved grandmother Firangiz who passed away during my academic job market phase in late 2020. I hope I made you proud.

I thank Refet Gürkaynak for helping me shine on.

Рукописи не горят.

Mikhail Bulgakov

Contents

Chapter 1

A Macroeconomic Model with Heterogeneous Banks

This chapter studies 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.

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.

1.1 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.

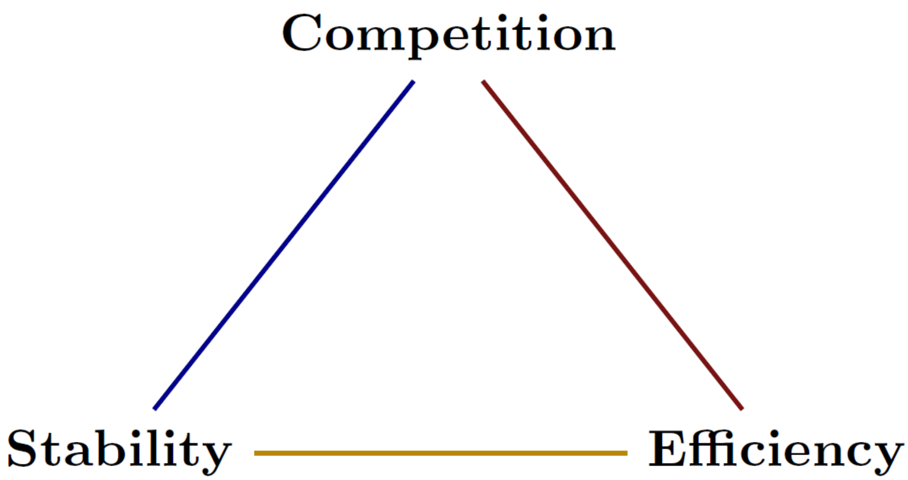
To formalize these facts into a coherent framework, I build a parsimonious dynamic general equilibrium macroeconomic model with heterogeneous banks. There are four main buildings blocks to this quantitative theory. First, we start with a stripped-down version of the workhorse representative-bank macroeconomic 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

[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 1.1: **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.

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 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 1.1 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

[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).

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; 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 Krish-

[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).

namurthy (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 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 1.2 provides three motivating facts on the cross section of U.S. banks. Section 1.3 lays out the model. Section 1.4 discusses how we take the model to the data. Section 1.5 presents the main quantitative analysis of the paper. Section 1.6 concludes. Finally, the Online Appendix contains additional applications, derivations, extensions, numerical algorithms, and data description.

1.2 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.

1.2.1 Leverage

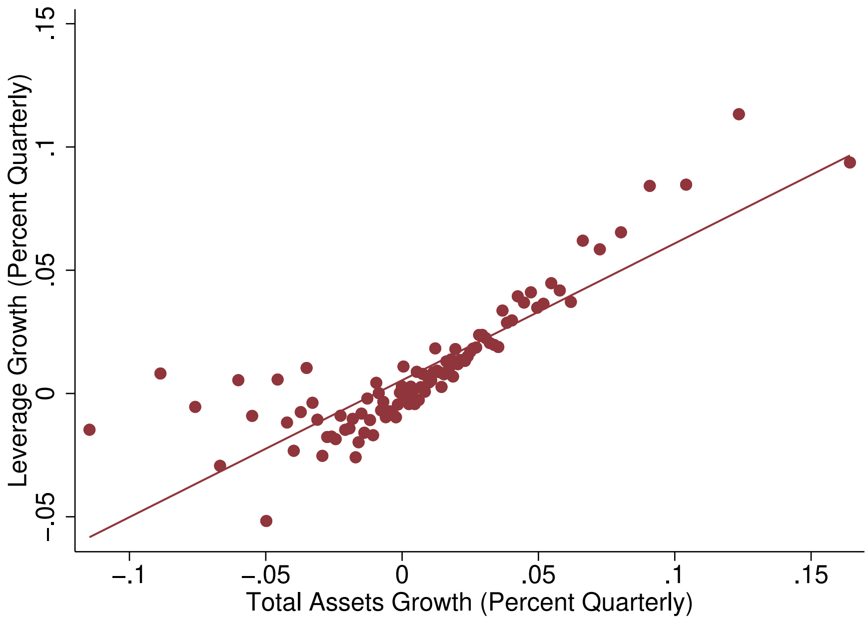
What is the relationship between bank leverage and balance sheet size? We answer this question using the data from Consolidated Reports of Condition and Income (the Call Reports). Every insured bank in the U.S. must submit these reports to the Federal Reserve on a quarterly basis. Following Adrian and 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 1.5 in Section 1.13 of the Online Appendix describes how every variable is defined and constructed.

Figure 1.2 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.

1.2.2 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

Figure 1.2: **Bank Assets and Leverage**

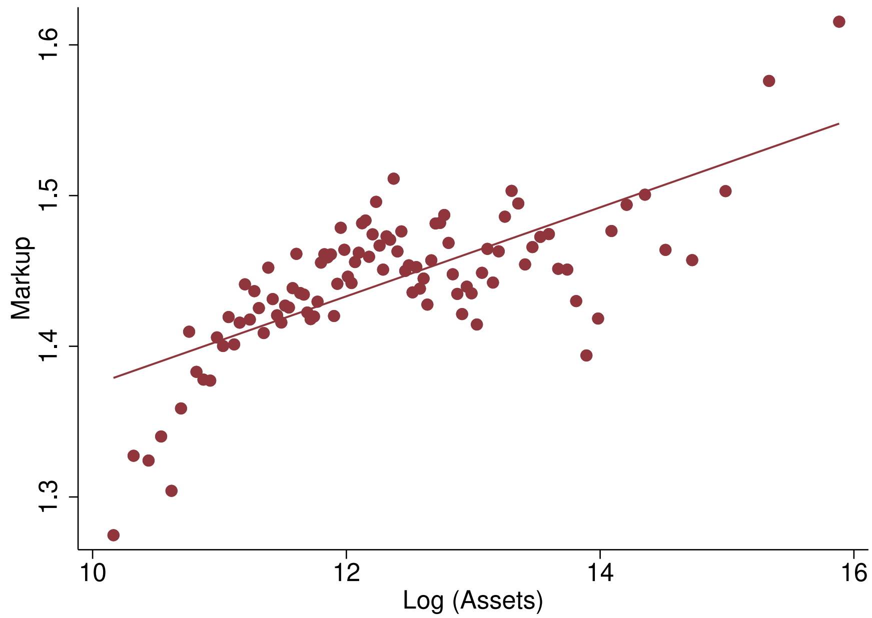


Notes: Binned scatter plot of bank assets and leverage growth in the data. The x-axis is binned into 100 percentiles of the distribution of quarterly assets growth. The y-axis is the bin-specific average of quarterly leverage growth. Leverage is defined as book assets over book equity. The leverage and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

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 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 1.3 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.

Figure 1.3: **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.

1.2.3 Exit Risk

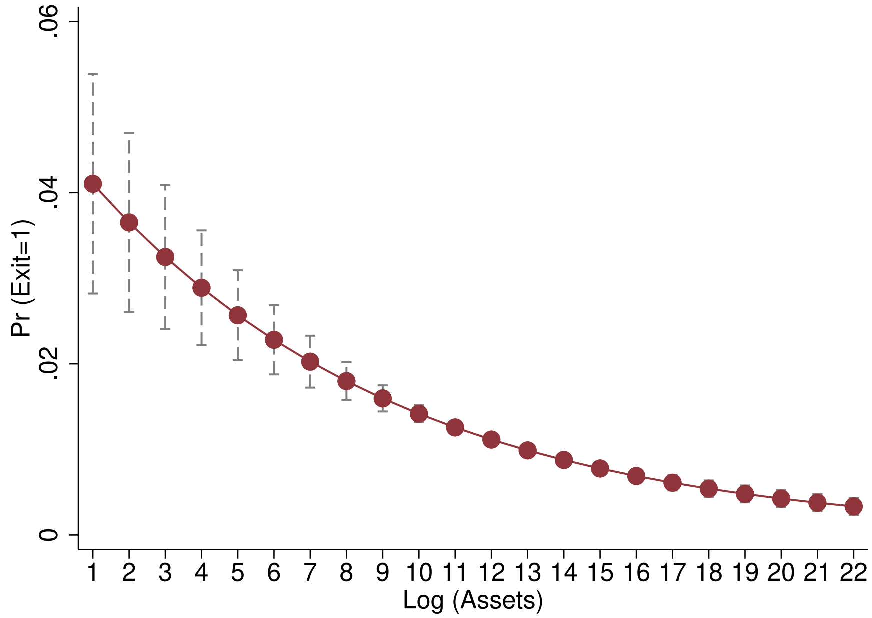
What is the relationship between bank balance sheet size and exit risk? I impute bank exit risk using two separate approaches.[6] The first approach relies on bank balance sheet and income statement data from the Call Reports. We begin by constructing an indicator variable Exit_{it} for each bank *i* 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 Exit_{it} on assets.

$$\Pr\{\text{Exit}_{\text{it}} = 1\} = \alpha_0 + \alpha_1 \text{Log (Assets)}_{\text{it}} + \mu_{\text{t}} + \epsilon_{\text{it}} \tag{1.1}$$

Where μ_{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 1.4 plots the margins plot from our logit regression and depicts all the point and interval estimates of the

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

Figure 1.4: **Bank Assets and Exit Risk**



Notes: Marginal point and interval estimates from the logit regression of the indicator for bank exit 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.

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 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.

1.3 A Model with Heterogeneous Intermediaries

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

1.3.1 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 1.5.2.

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 1.7.2 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 1.11.3 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.

1.3.2 Technology

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

$$Y_t = AK_t^{\alpha}L_t^{1-\alpha} \tag{1.2}$$

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:

$$R_{t+1}^k = \frac{A\alpha K_{t+1}^{\alpha-1}}{P_t} \qquad W_t = (1-\alpha)AK_t^{\alpha} \tag{1.3}$$

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:

$$K_t = \int_0^1 \Upsilon\left(\frac{k_t(j)}{K_t}\right) dj \tag{1.4}$$

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:

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

subject to technology 3.4. This yields a demand function for bank funds:

$$p_t(j) = \Upsilon'\left(\frac{k_t(j)}{K_t}\right) Z_t \tag{1.5}$$

where

$$Z_t := \left(\int_0^1 \Upsilon'\left(\frac{k_t(j)}{K_t}\right)\frac{k_t(j)}{K_t} dj\right)^{-1} \tag{1.6}$$

is the *Kimball demand index*. In the Dixit-Stiglitz special case, $Z_t = \frac{\theta}{\theta-1}$, and (1.5) 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 θ. Section 1.11.1 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.

1.3.3 Banks

The general credit demand system in (3.4)-(1.6) 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:

$$d_t(j) + n_t(j) = p_t(j)k_t(j) \tag{1.7}$$

Every bank faces non-interest expenses $\frac{1}{\zeta_1}k_t(j)^{\zeta_2}$ where parameter ζ_2 can help govern the degree of non-linearity and scale-variance. Section 1.10.1 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:

$$R_t^T(j) = \kappa\xi_t(j) + (1-\kappa)R_t^k \tag{1.8}$$

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 1.11.2 of the Online Appendix. The idiosyncratic return, $\xi \in \Xi$, follows an AR(1) process:

$$\xi_t(j) = (1-\rho_\xi)\mu_\xi + \rho_\xi\xi_{t-1}(j) + \sigma_\xi\epsilon_t(j) \tag{1.9}$$

Analogously, let the finite state Markov representation of (3.10) be $\mathbf{G}(\xi_{t+1}, \xi_t)$. We can now state the law of motion of bank-level net worth:

$$n_{t+1}(j) = R_{t+1}^T(j)p_t(j)k_t(j) - \bar{R}_t(j)d_t(j) - \frac{1}{\zeta_1}k_t(j)^{\zeta_2} \tag{1.10}$$

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 λ of the total value of revenues p(j)k(j). If deciding to divert, the banker always escapes but the franchise enters bankruptcy the following period. The banker is indifferent between operating honestly and diverting when whatever he is able to divert exactly equals the value of the franchise. This yields the following incentive constraint that puts a limit on bank leverage.

$$\lambda p_t(j)k_t(j) \leq V_t(j) \tag{1.11}$$

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.:

$$\upsilon_t(j) = \Pr\Big(n_{t+1}(j) \leq 0\Big) \tag{1.12}$$

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 1.2.

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 ξ. The law of motion for the distribution is:

$$\eta_{t+1}(n_{t+1}, \xi_{t+1}) = \Phi(\eta_t) \tag{1.13}$$

We define Φ in detail below.

Dynamic Problem of the Incumbent Banker The following summarizes the dynamic problem of the incumbent. We adopt recursive notation because the solution does not depend on a specific bank j but on the relevant state variables only. Define $\mathbf{s} = \{n, \xi\}$ as the bank's idiosyncratic state vector. There is no aggregate risk. The bank maximizes its franchise value which is defined as the discounted stream of future flows of net worth. With an exogenous probability σ the incumbent may exit involuntarily, in which case 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 Λ, 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 η and its law of motion Φ, bank-specific deposit rates $\bar{R}$ and portfolio returns R^T. Each bank solves:

$$V(\mathbf{s}) = \max_{\{k,p,d\}} \left\{ \mathbb{E}_{\mathbf{s'}|\mathbf{s}} \left[\Lambda' \Big((1-\sigma) n' + \sigma V(\mathbf{s'}) \Big) \right] \right\} \tag{1.14}$$

s.t. conditions 3.4-3.14.

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:

$$\max_{\phi} \left[\mu_a \phi + \nu_a \right] \tag{1.15}$$

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 1.10.2 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)**.**

$$\frac{p(j)}{P} = \mu(x)\frac{k(j)^{\zeta_2 - 1}}{R^T(j) - \bar{R}(j)} \tag{1.16}$$

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.

1.3.4 Variable Markups

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

$$\Upsilon(x) = 1 + (\theta - 1)\exp\frac{1}{\epsilon}\epsilon^{\frac{\theta}{\epsilon}-1}\left[\Gamma\left(\frac{\theta}{\epsilon}, \frac{1}{\epsilon}\right) - \Gamma\left(\frac{\theta}{\epsilon}, \frac{x^{\frac{\epsilon}{\theta}}}{\epsilon}\right)\right] \tag{1.17}$$

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:

$$\Gamma(s, q) := \int_q^\infty t^{s-1}\exp^{-t} dt \tag{1.18}$$

The CES aggregator is a special case of (1.17) when $\epsilon = 0$. With the Klenow-Willis specification, we have:

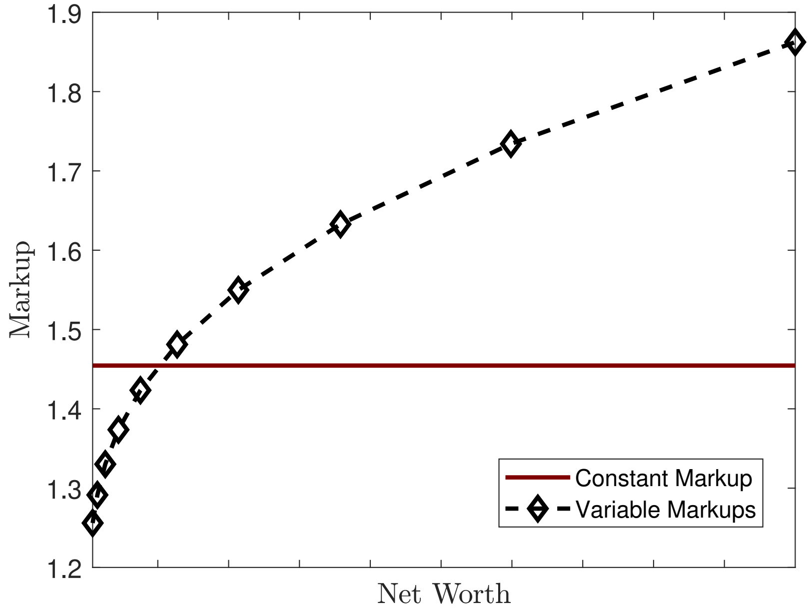
$$\Upsilon'(x) = \frac{\theta - 1}{\theta}\left(\exp\frac{1 - x^{\frac{\epsilon}{\theta}}}{\epsilon}\right) \tag{1.19}$$

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:

$$\mu(x) = \frac{\theta x^{-\frac{\epsilon}{\theta}}}{\theta x^{-\frac{\epsilon}{\theta}} - 1} \tag{1.20}$$

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 1.5: **Bank Markups**



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

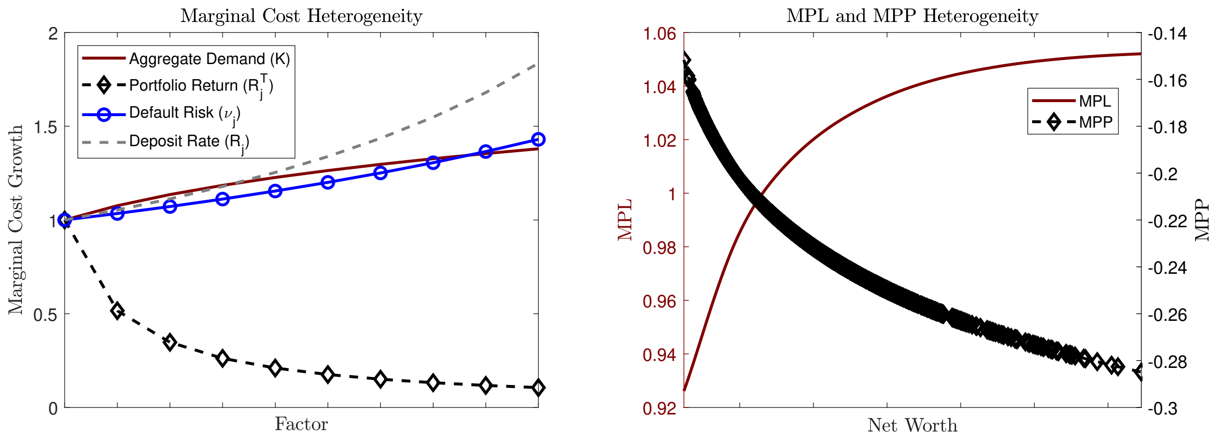
can be achieved in a simple way by varying $\frac{\epsilon}{\theta}$. When taking the model to the data, we will use the empirical cross-section of bank markups to deduce the two parameters.

Figure 1.5 illustrates the differences between Kimball-Klenow-Willis and Dixit-Stiglitz aggregators. Increasing ϵ 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 1.2. We will discuss the calibration strategy that achieves this in Section 1.4.

1.3.5 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 $\mathrm{R}^{\mathrm{T}}(\mathrm{j})$, interest rate on deposits $\bar{\mathrm{R}}(\mathrm{j})$, the scale effect in K, and the probability of default $\nu(\mathrm{j})$ (which affects marginal costs indirectly, through $\bar{\mathrm{R}}(\mathrm{j})$). 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

Figure 1.6: **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.

cost is summarized on the left panel of Figure 1.6. First, we observe that bank-level marginal 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):

$$\mathrm{MPL} = \int_{\mathbf{B}} \frac{\partial \mathrm{k(j)}}{\partial \mathrm{n(j)}} \eta(\mathrm{dn}, \mathrm{d}\xi) \quad \mathrm{MPP} = \int_{\mathbf{B}} \frac{\partial \mathrm{p(j)}}{\partial \mathrm{n(j)}} \eta(\mathrm{dn}, \mathrm{d}\xi) \tag{1.21}$$

The right panel of Figure 1.6 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.

1.3.6 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 μ_ξ. As mentioned before, we relax the assumption of exogenous entry in Section 1.7.2 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 σ 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.

1.3.7 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:

$$\eta'(n',\xi') = \sum_{\xi} G(\xi',\xi) \int \mathbb{1}_{\{(n,\xi)|K(n,\xi)\in\mathbf{B}\}} \times \mathbb{1}_{\{\bar{d}(n,\xi)=0\}} \eta(dn, d\xi) + M' n_0 \tag{1.22}$$

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 ξ of the incumbents.

1.3.8 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:

[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).

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

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

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

$$\bar{R}_t(j) = \frac{1 - \upsilon_t(j) x_t(j) \mathbb{E}\left(R^T_{t+1}(j)\Lambda_{t+1}\right)}{\left(1 - \upsilon_t(j)\right)\mathbb{E}\left(\Lambda_{t+1}\right)} \tag{1.23}$$

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 ϕ the market leverage ratio, defined as before.

1.3.9 Stationary Industry Equilibrium

Credit market clearing requires:

$$K = \int_{\mathbf{B}} \Big(k(n, \xi)\Big)\eta(dn, d\xi) \tag{1.24}$$

Similarly, clearing the deposit market requires:

$$\int_0^1 b(j)dj = \int_{\mathbf{B}} \Big(d(n, \xi)\Big)\eta(dn, d\xi) \tag{1.25}$$

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 Λ, 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 (3.27). The deposit market clears as in (3.28)
6. The cross-sectional distribution evolves according to (3.25) and is consistent with bank-level demand functions

1.3.10 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: Λ, 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 1.12 of the Online Appendix.

1.4 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.

1.4.1 Parameterization

All chosen parameters are shown in Table 1.1. The model period is one quarter. We begin by describing standard macro parameters. We set the share of aggregate capital in production α to

Table 1.1: **Parameter Values**

Parameter	Description	Value
	Macro	
α	Share of capital in production	0.36
β	Discount factor	0.996
σ_{h}	Risk aversion	1
	Banking	
σ	Dividend payout ratio	0.9
ω	Share of divertible assets	0.1
n_0	New banker endowment	30% of N
$\frac{1}{\zeta_1}$	Monitoring cost linear	0.01
ζ_2	Monitoring cost quadratic	1.19
	Monopolistic Credit Market	
θ	CES elasticity	3.2
$\frac{\epsilon}{\theta}$	Superelasticity	0.165
	Idiosyncratic Bank Return Risk	
κ	Fraction of portfolio exposed to idiosyncratic risk	0.3
ρ_ξ	Serial correlation of idiosyncratic risk	0.52
σ_ξ	SD of idiosyncratic risk	0.085
	Costly Bank Default	
d_1	Default cost constant	0.0511
d_2	Default cost linear	0.0075

0.36. The discount factor β 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 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 (ζ_1, ζ_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]

[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 1.10.1.

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. Second, the superelasticity $\frac{\epsilon}{\theta} = 0.165$ generates the variable markup function seen on Figure 1.5.[16]

Parameters from the idiosyncratic return shocks block are chosen in order to match three facts. First, we motivate κ 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, σ_ξ 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 1.2.3. Third, persistence ρ_ξ 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 1.2.3. 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:

$$\text{Default Cost(j)} = \mathrm{d}_1 + \mathrm{d}_2\mathrm{k(j)} \tag{1.26}$$

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.

1.4.2 Validation

[15]Using an approach that does not rely on production functions, Jamilov (2020) estimates branch-level loan price 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]ρ_ξ is one of the key parameters in the model that directly impacts concentration in the banking sector. A large ρ_ξ (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 ρ_ξ 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 1.2: **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

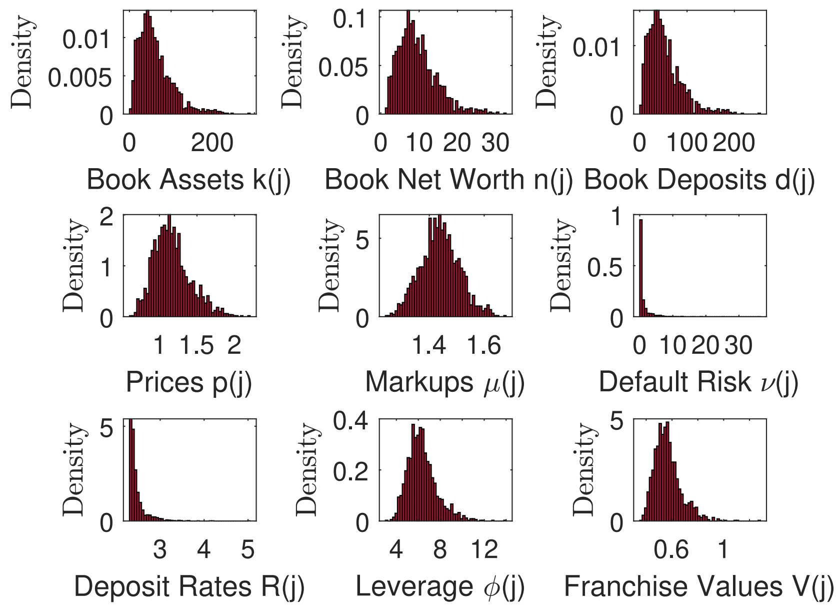
Notes: key moments in the model and in the data. Variables are defined in Table 1.5. Probabilities are annualized. Markups are absolute. Data source: Call Reports. Quarterly data is over 2010:q1-2019:q4.

Moments Table 1.2 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 1.5 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 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.

Figure 1.7: **Stationary Distributions in the Model**

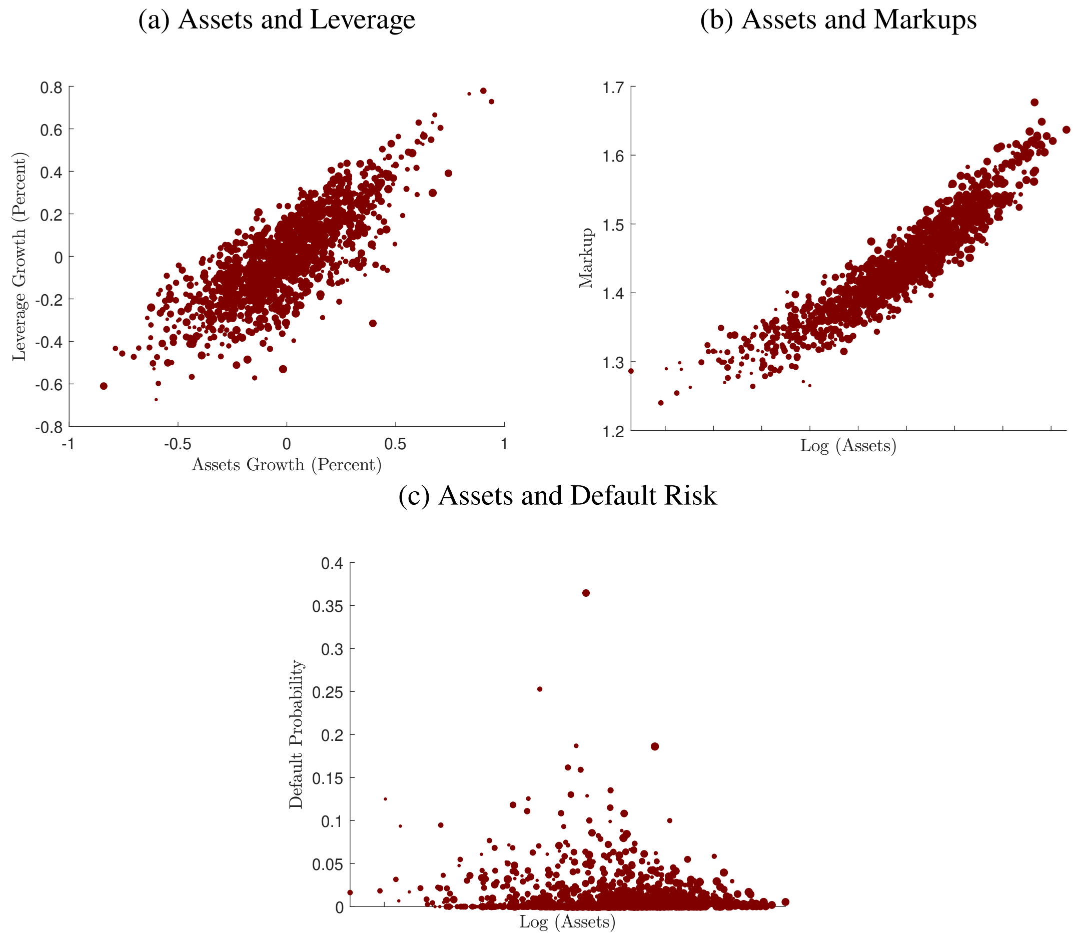


Notes: Model-generated stationary distributions.

Distributions We now look at all stationary cross-sectional distributions that the model generates. Figure 1.7 plots univariate histograms for bank assets k(j), net worth n(j), deposits d(j), market leverage ϕ(j) relative prices p(j), markups μ(j), default risk ν(j), deposit rates $\bar{\text{R}}$(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 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 1.5. First, in line with the data, the model generates a positive cross-sectional correlation between bank leverage and assets growth. Figure 1.8 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

Figure 1.8: **Cross-Sectional Patterns in the Model**



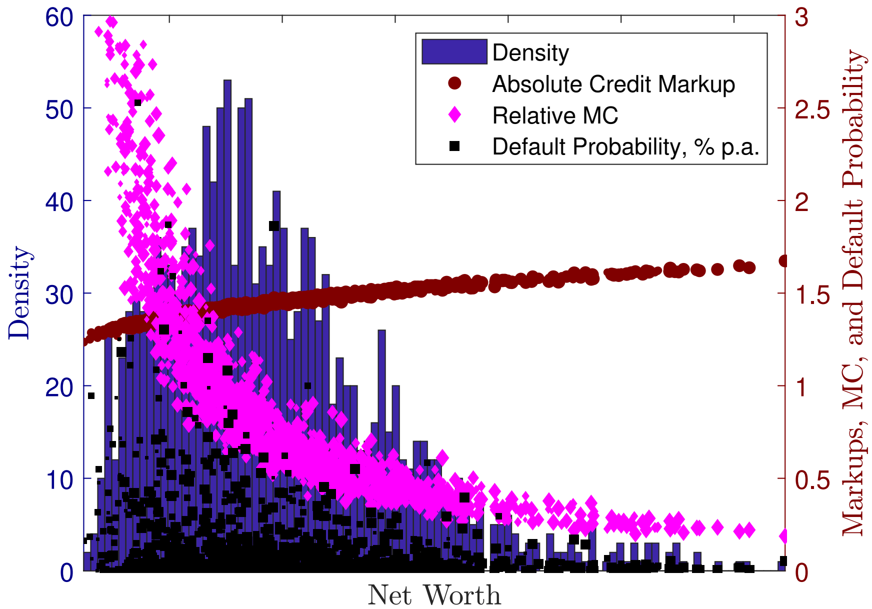
Notes: Cross-sectional relationships between measures of bank size, default risk, book leverage, and markups in the model. All panels show scatter plots based on a stochastic simulation of the baseline model with N=1 intermediaries and T=2,000 quarters.

from the scatter plot.

Second, the model correctly predicts that larger banks choose to charge higher markups μ(j). This can be seen clearly from Panel (b) of Figure 1.8. 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 model predicts that the cross section of default risk ν(j) is concentrated in the left tail of the assets distribution. Figure 1.8, 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 1.9: **The Banking Industry Trilemma**



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

1.5 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]

1.5.1 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 1.9 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 1.3 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 $\nu(\mathrm{j})$ 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.

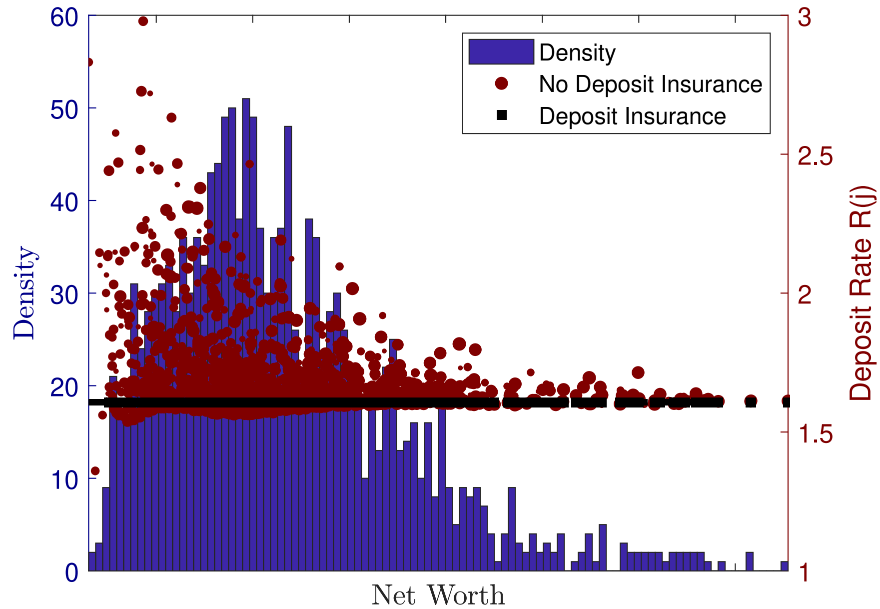
It is crucial to emphasize that the trilemma does not imply that it is *impossible* for the regulator to improve *net* welfare. Net quantitative effects always depend on calibration of the credit demand superelasticity, the cost of default function, and of the idiosyncratic risk process. Suppose we consider 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.

1.5.2 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 $\bar{\mathrm{R}}(\mathrm{j})$ is flat. To achieve this as an endogenous result, banks continue to take as given their individual default probability $\nu(\mathrm{j})$, but we break the mapping between $\nu(\mathrm{j})$, the deposit recovery rates $\mathrm{x}(\mathrm{j})$, and rates $\bar{\mathrm{R}}(\mathrm{j})$. 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 1.10: **Deposit Insurance Scheme**



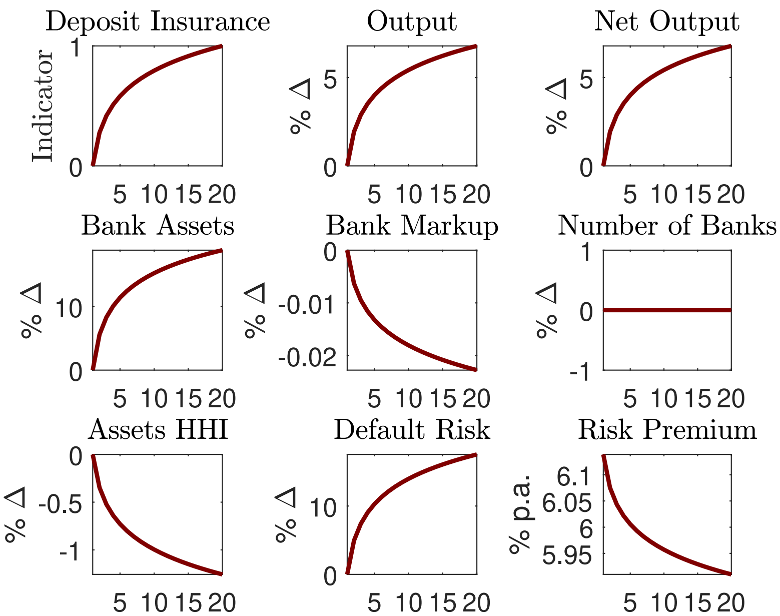
Notes: distribution of bank net worth and scatter plots for deposit rates in the economy with deposit insurance. Black squares represent equilibrium rates when guarantees are turned on and red circles the counterfactual rates if guarantees were turned off.

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 1.10 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 1.11 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 1.11: **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.

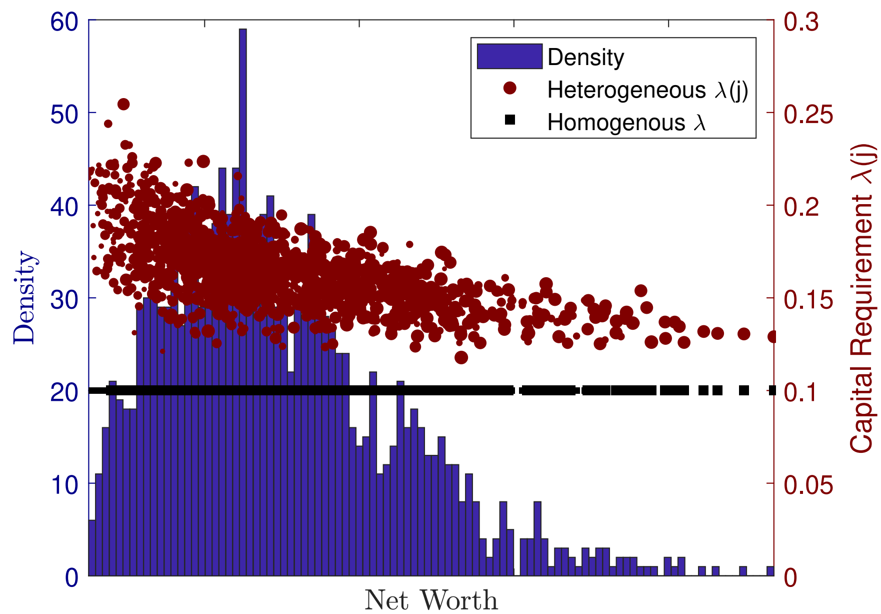
we see a positive effect on average default risk. Since the deposit insurance scheme favors small banks by more, the market share of of low-net-worth banks increases and concentration (Assets HHI) falls. As a result, the average markup falls. Finally, aggregate risk premium, defined as $R^k - \bar{R}$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

1.5.3 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 λ 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 1.12: **Heterogeneous Capital Requirements**



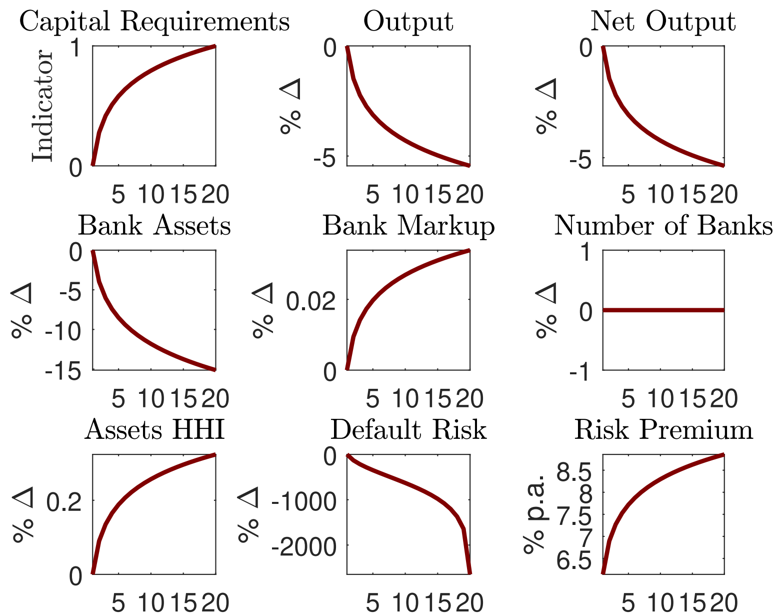
Notes: distribution of bank net worth and scatter plots for λ under homogenous and heterogeneous capital requirement regimes.

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 1.12 shows how the policy works in the model. Overlayed on the new equilibrium distribution of net worth are the homogenous λ 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 1.26.

Figure 1.13 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 1.13: **Macroeconomic Effects of Heterogeneous Capital Requirements**



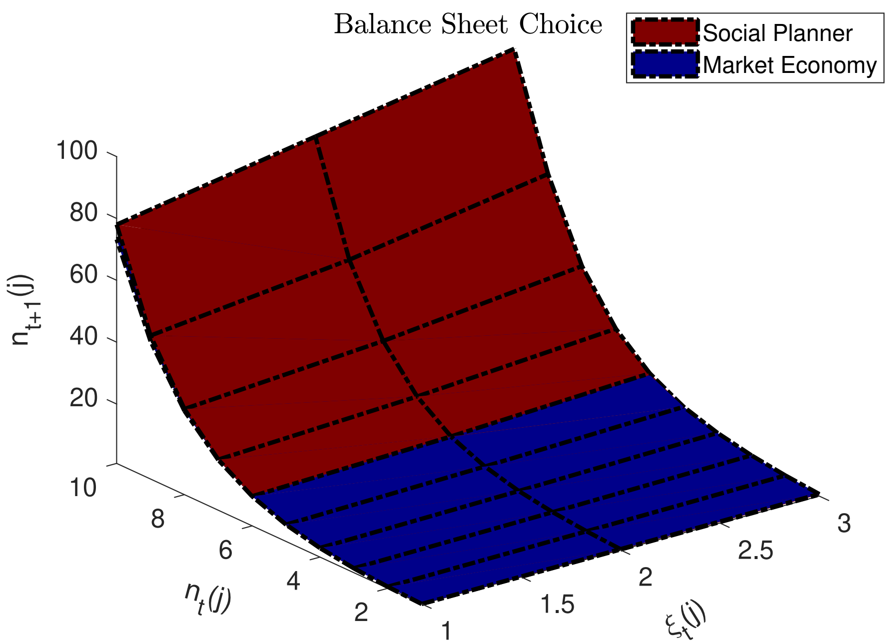
Notes: macroeconomic effects of switching on heterogeneous capital requirements.

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 $\lambda(\mathrm{j})$ 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.

1.5.4 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 $\{\mathrm{k}, \mathrm{d}, \mathrm{p}, \mu\}$ 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 1.14: **Market Equilibrium and Constrained Efficient Allocations**



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

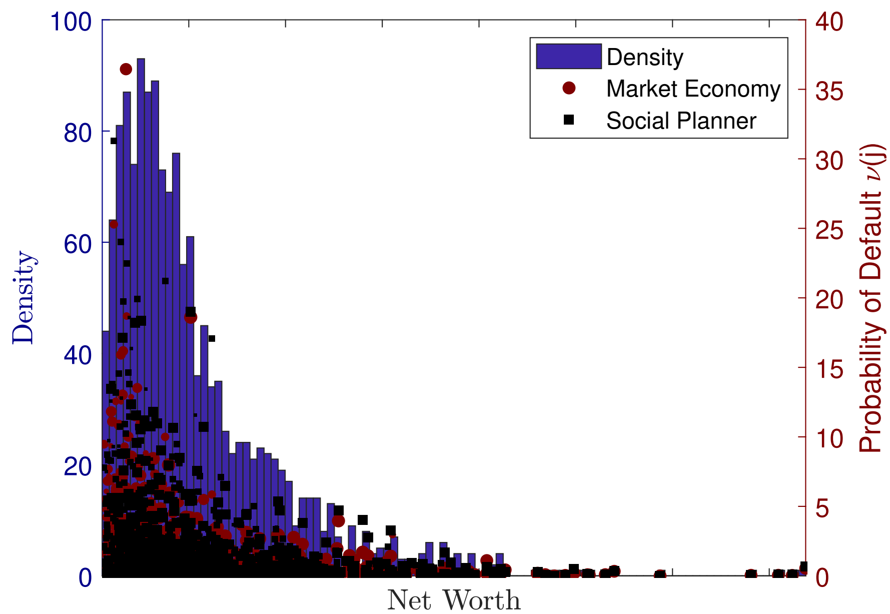
$$n_{t+1}(j) = R^T\Big(n(j), \xi(j), \{k_t(j), d_t(j), p_t(j)\}\Big)p_t(j)k_t(j) - \bar{R}_t(j)d_t(j) - \frac{1}{\zeta_1}k_t(j)^{\zeta_2} \tag{1.27}$$

Compare this formula to Equation 3.11 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 1.12 of the Online Appendix for more details on the numerical algorithm.

Figure 1.14 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 1.15: **Systemic Risk Implications of Constrained Efficiency**



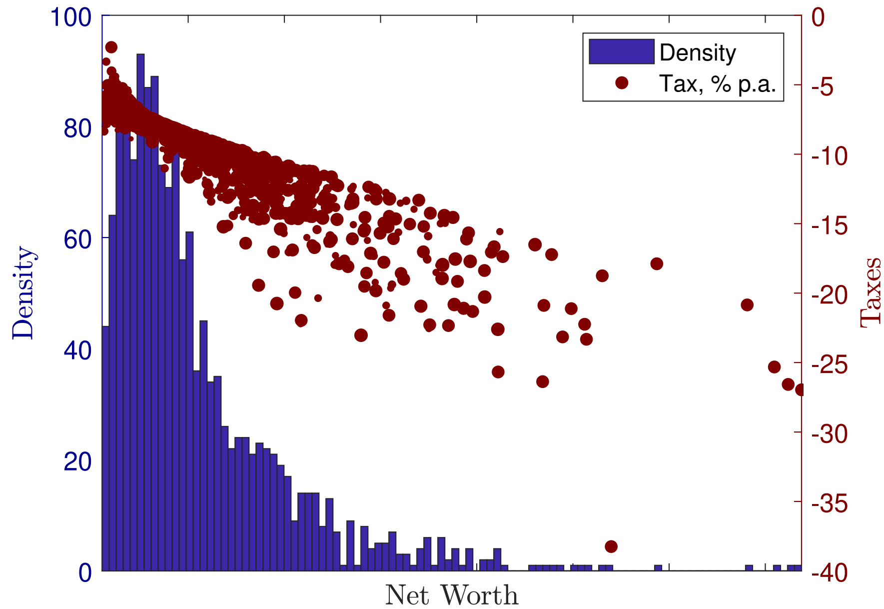
Notes: distribution of bank net worth and scatter plots for ν(j) under alternative market regimes.

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

Figure 1.15 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 ν(j). We plot the new stationary distribution of net worth and ν(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 1.16: **Optimal Policy**



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

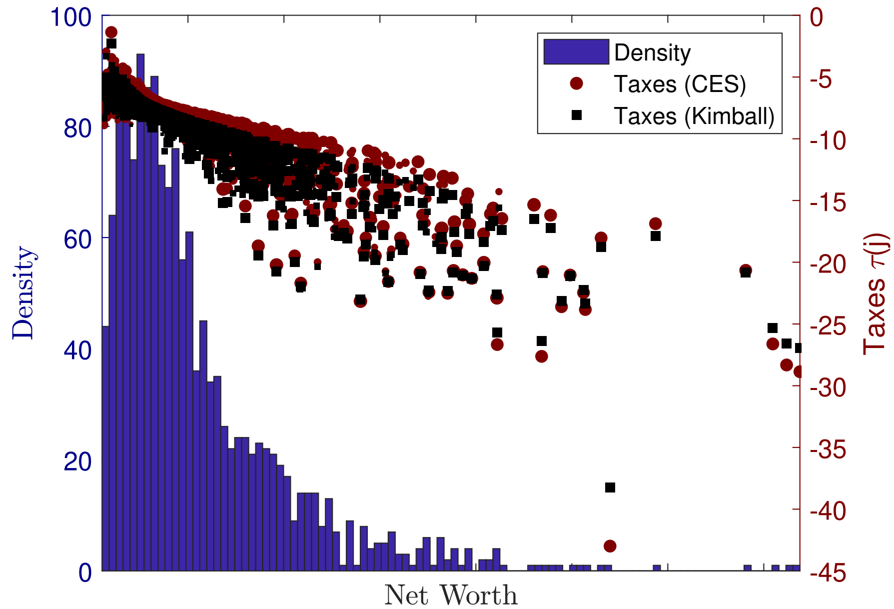
bank-specific total portfolio return process and the law of motion of net worth. Specifically, we conjecture a size and idiosyncratic return specific tax rule $\tau\Big(\mathrm{n(j)}, \xi(\mathrm{j})\Big)$ and impose it on the market equilibrium. Computationally, we assume that taxes are polynomials in n(j) and $\xi(\mathrm{j})$ 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:

$$\mathrm{n_{t+1}(j)} = \mathrm{R_t^T(j)}\left[1 - \tau\Big(\mathrm{n(j)}, \xi(\mathrm{j})\Big)\right]\mathrm{p_t(j)k_t(j)} - \bar{\mathrm{R}}_\mathrm{t}\mathrm{(j)d_t(j)} - \frac{1}{\zeta_1}\mathrm{k_t(j)}^{\zeta_2} \tag{1.28}$$

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

Figure 1.16 plots the stationary distribution of net worth from the social planner's problem with the scatter plot for optimal taxes $\tau(\mathrm{j})$. 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 1.17: **Optimal Taxes under CES and Kimball Aggregators**

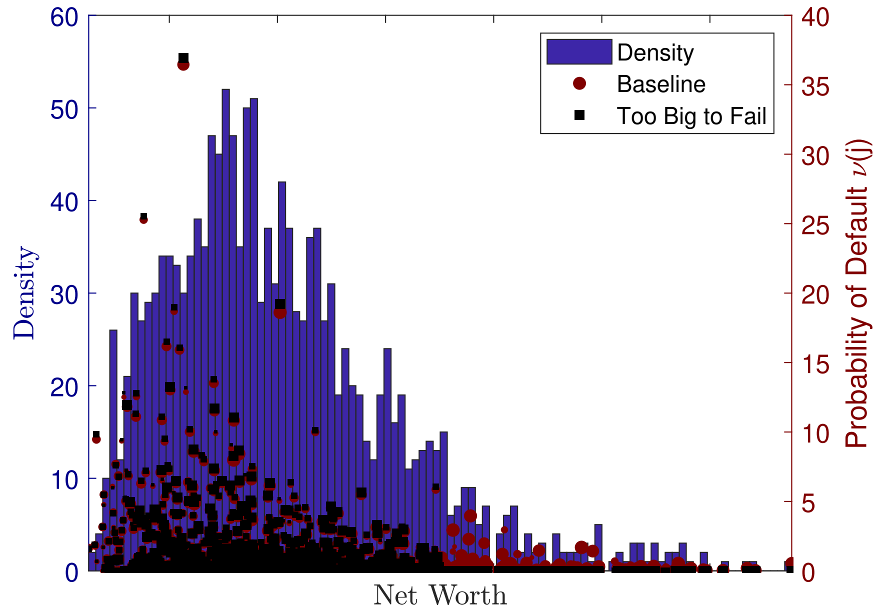


Notes: optimal taxes τ(j) under constant (CES) and variable (Kimball) markups.

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 1.17 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 1.18: **Too Big to Fail**



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

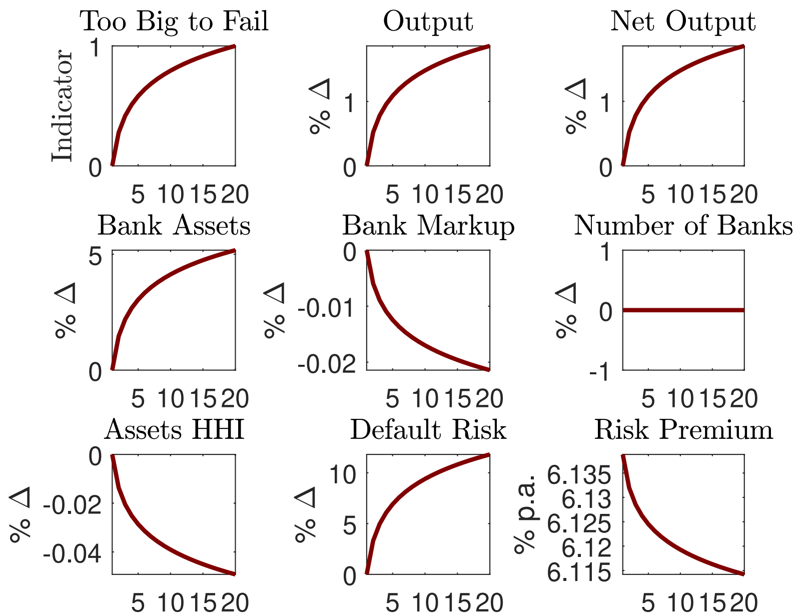
1.5.5 Too Big to Fail

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 ν(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 ν(j) that is consistent with their size-risk profile as usual. The policy function for ν(j) is therefore kinked, and all banks understand this. The pass-through from ν(j) to $\bar{\text{R}}$(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 1.19: **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 1.18 shows how the mechanism works. Observe how the scatter plot for ν(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 ν(j). In addition, it is interesting that most banks in the bottom nine deciles now face a *higher* ν(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 1.19 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 1.3: **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.

1.5.6 Summary of All Allocations

To summarize our findings across different regulatory regimes and market structures, we report all key aggregates in Table 1.3. 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.

1.5.7 Additional Results and Applications

The Online Appendix provides further results and quantitative applications. In Section 1.7 I analyze applications of the model to the rise of banking concentration, emergence of fintech-intermediated credit, and intermediary asset pricing. Section 1.8 studies targeted bank-level stabilization policies such as equity injections and liquidity facilities. Section 1.9 simulates MIT shocks to aggregate productivity.

1.6 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.

Appendix

1.7 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.

1.7.1 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 $k_1 \ldots k_N$ 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.

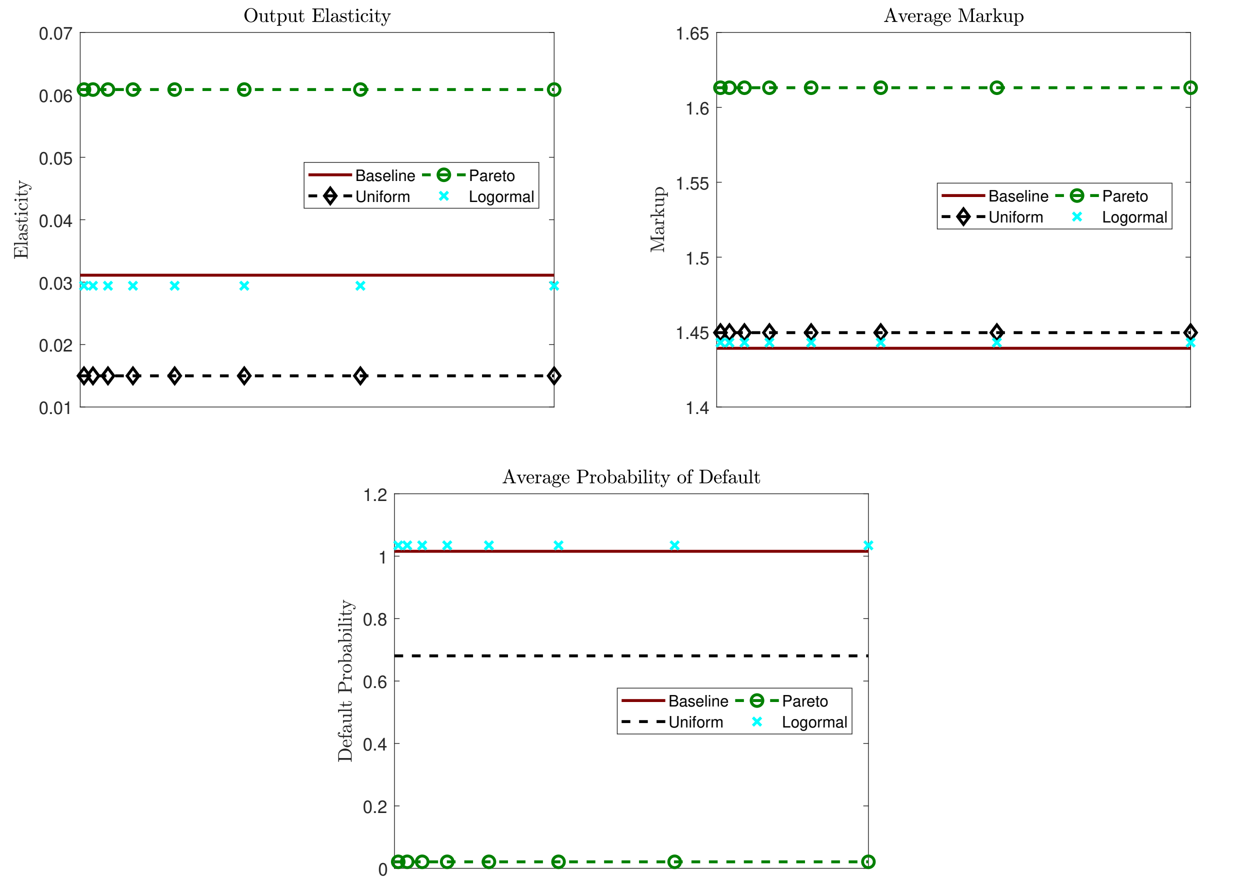
We consider 3 broad families of densities: Uniform, Lognormal, and Pareto. For the uniform density, we generate N=2,000 random numbers from the interval $[0.5K_{ss}, 1.5K_{ss}]$, where K_{ss} stands for the level of aggregate capital in the market equilibrium. For the lognormal density, we draw from $P(\mu_k, \sigma_k^2)$, where μ_k and σ_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 $\alpha = 2$.[23]

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):

$$\frac{\partial Y}{\partial N} = \underbrace{\frac{\partial Y}{\partial K}}_{\text{MPK}} \times \underbrace{\int_{\mathbf{B}} \frac{\partial k(j)}{\partial n(j)} \mu(dn, d\xi)}_{\text{MPL(j)}} \tag{1.29}$$

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

Figure 1.20: **Macroeconomic Effects of Alternative Banking Distributions**



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

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.

Figure 1.20 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 1.9, 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 1.20 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.

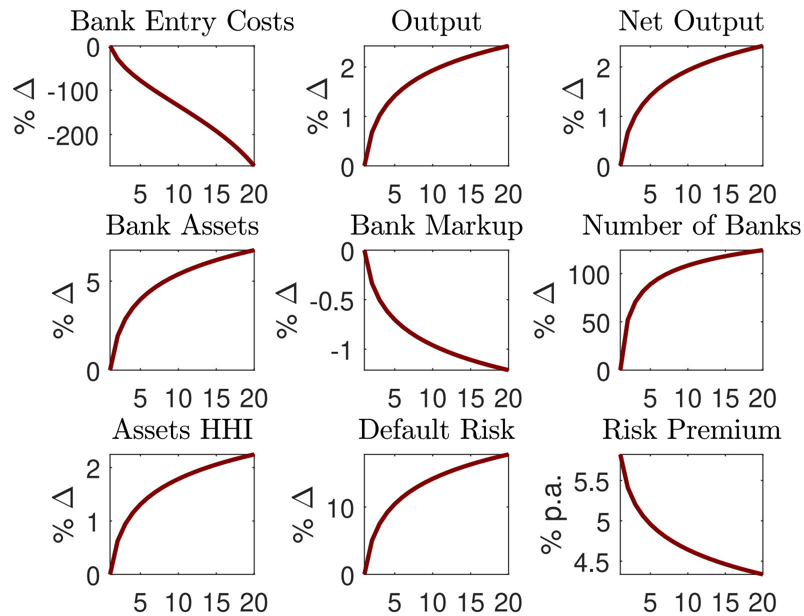
Overall, this exercise is a simple but useful demonstration of quantitative implications of the banking trilemma. The most concentrated economy, whose distribution is drawn from a Pareto density, is the most efficient, least competitive, and most stable. The prediction of our framework for the future 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).

1.7.2 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 $\mathrm{G}_0(\xi)$ that is implied by the ξ process. The financier is also bestowed with an initial level of net worth n_0 which is a constant fraction of the aggregate stock of net worth N. Afterwards, the financier decides whether to operate or to immediately exit. Conditional on its state $\{\mathrm{n}_0, \xi_0\}$, the financier operates if and only if its expected discounted franchise value exceeds e. The value function of the entering financier is therefore:

$$\mathrm{V}^{\mathrm{e}}(\mathrm{n}_0, \xi_0) \equiv \max\left[\mathrm{V}(\mathrm{n}_0, \xi_0) - e, 0\right] \tag{1.30}$$

Figure 1.21: **Fintech Credit Growth**



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

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

The 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:

$$\mu'(n', \xi') = \underbrace{\sum_{\xi} G(\xi', \xi) \int \mathbb{1}_{\{(n,\xi)|K(n,\xi)\in\mathbf{B}\}} \times \mathbb{1}_{\{\bar{d}(n,\xi)=0\}} \eta(dn, d\xi)}_{\text{Incumbents}} + \underbrace{M' \int \mathbb{1}_{\{(n_0,\xi)|K(n,\xi)\in\mathbf{B}\}} G_0(\xi)}_{\text{New Entrants}} \tag{1.31}$$

The law of motion of the distribution is now:

$$\eta_{t+1}(n_{t+1}, \xi_{t+1}) = \Phi(\eta_t, M_{t+1}) \tag{1.32}$$

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

$$\underbrace{\mathrm{K}}_{\text{Aggregate Supply}} = \underbrace{\int_{\mathbf{B}} \Big(\mathrm{k}(\mathrm{n}, \xi)\Big) \eta(\mathrm{dn}, \mathrm{d}\xi)}_{\text{Incumbent Demand}} + \underbrace{\mathrm{M} \int_{\mathbf{B}} \Big(\mathrm{k}(\mathrm{n}_0, \xi_0)\Big) \mathrm{dG}(\xi_0)}_{\text{Entrants Demand}} + \underbrace{\mathrm{M}e}_{\text{Entry Cost}} \tag{1.33}$$

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 1.21 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.

1.7.3 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:

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

Where $\varphi(\mathrm{j})$ 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

Table 1.4: **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.

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.

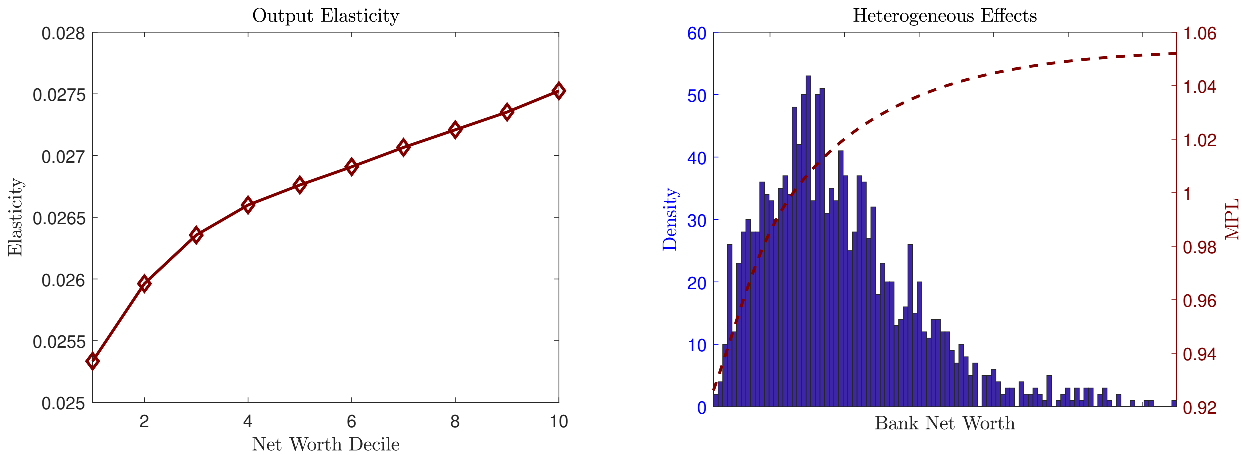
Table 1.4 presents key asset pricing moments from the framework under different assumptions. Without any heterogeneity, the liquidity and default risk channels generate a premium of 2.28%. Adding monopolistic competition with variable markups and uninsurable idiosyncratic shocks gets us a large risk 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.

1.8 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.[24]

[24] 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 1.5.2 which was an aggregate policy. Results for targeted direct lending and bank-level debt guarantees are available upon request.

Figure 1.22: **Macroeconomic Effects of Targeted Equity Injections**



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

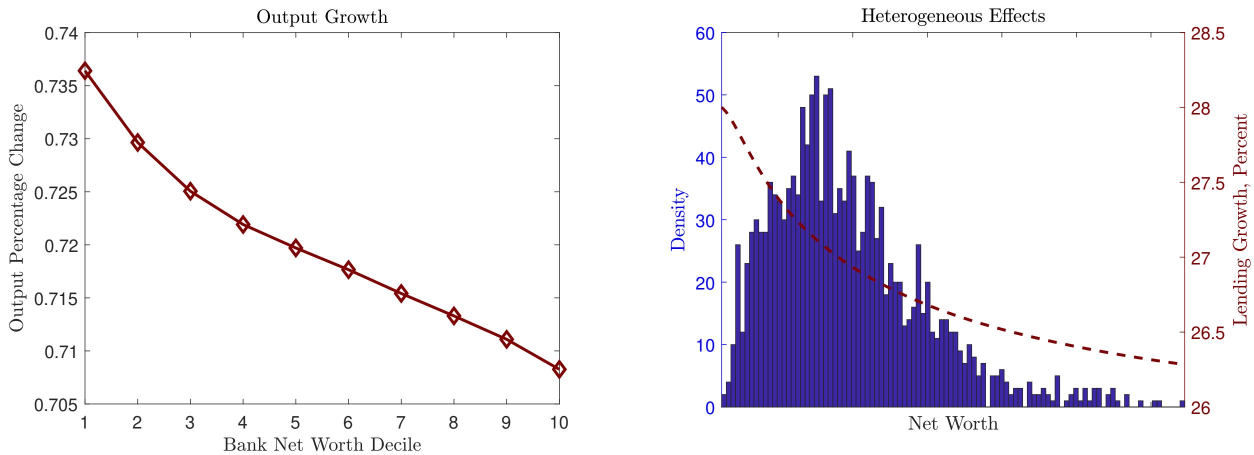
1.8.1 Equity Injections

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 $\iota = 1 \dots 10$, we assume that the government increases by one percent the net worth of each bank in ι but not anywhere else in the 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 1.22 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.[25]

[25]These bailouts are unexpected and do not generate additional moral hazard frictions ex-ante. The implicit bailout subsidy is internalized in an earlier Section 1.5.5.

Figure 1.23: **Macroeconomic Effects of Targeted Liquidity Facilities**



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

1.8.2 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 - λ - controls the degree of constraint tightness and is generally part of the exogenous environment. We now suppose that the government can step in and augment λ on behalf of private lenders. In particular, we allow λ to be relaxed on *any* bank in the distribution. In practice, this intervention can be mapped to 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.[26] With the binding leverage constraint, it is straightforward to solve for the bank-specific leverage ratio:

$$\phi(j) = \frac{\nu_a(j)}{\lambda - \mu_a(j)} \tag{1.34}$$

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 λ) increases banks appetite for leverage. Everything else equal, this raises credit supply in the market.

We proceed by assuming that the government intervenes by lowering λ_ι on decile $\iota = 1 \ldots 10$ of the banks net worth distribution by 10% relative to the baseline value of 0.1. The exogenous shock

[26]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.

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 1.23 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 λ_ι on ϕ_ι is negative and declining with bank size due to diminishing marginal costs of funds.

1.9 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 $\{\Omega_t\} = \{R^k_t, \bar{R}_t, P_t, \tau_t\}$ and where τ_t summarizes any policy actions of the government:

$$K_t\Big(\{\Omega_t\}_{t\geq 0}\Big) = \int k_t\Big(n, \xi; A_t, \{\Omega_t\}_{t\geq 0}\Big)\mu_t(dn, d\xi) \tag{1.35}$$

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 1.35:[27]

$$dK_0 = \Big[\underbrace{\int_0^\infty \frac{\partial K_0}{\partial A_t} dA_t}_{\text{Direct Effect}} + \underbrace{\int_0^\infty \Big(\frac{\partial K_t}{\partial \bar{R}_t} d\bar{R}_t + \frac{\partial K_t}{\partial R^k_t} dR^k_t + \frac{\partial K_t}{\partial P_t} dP_t + \frac{\partial K_t}{\partial \tau_t} d\tau_t\Big)}_{\text{Indirect Effect}}\Big] dt \tag{1.36}$$

The first term in Equation 1.36 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

[27]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.

effect can be further decomposed into the cross section of bank-level marginal propensities to lend:

$$\int_0^\infty \frac{\partial K_0}{\partial A_t} dA_t dt = \int_0^\infty \Big[\int \frac{\partial k_0(n, \xi; A_t, \{\bar{\Omega}_t\}_{t\geq 0})}{\partial A_t} \bar{\mu}(dn, d\xi) \Big] dA_t dt \tag{1.37}$$

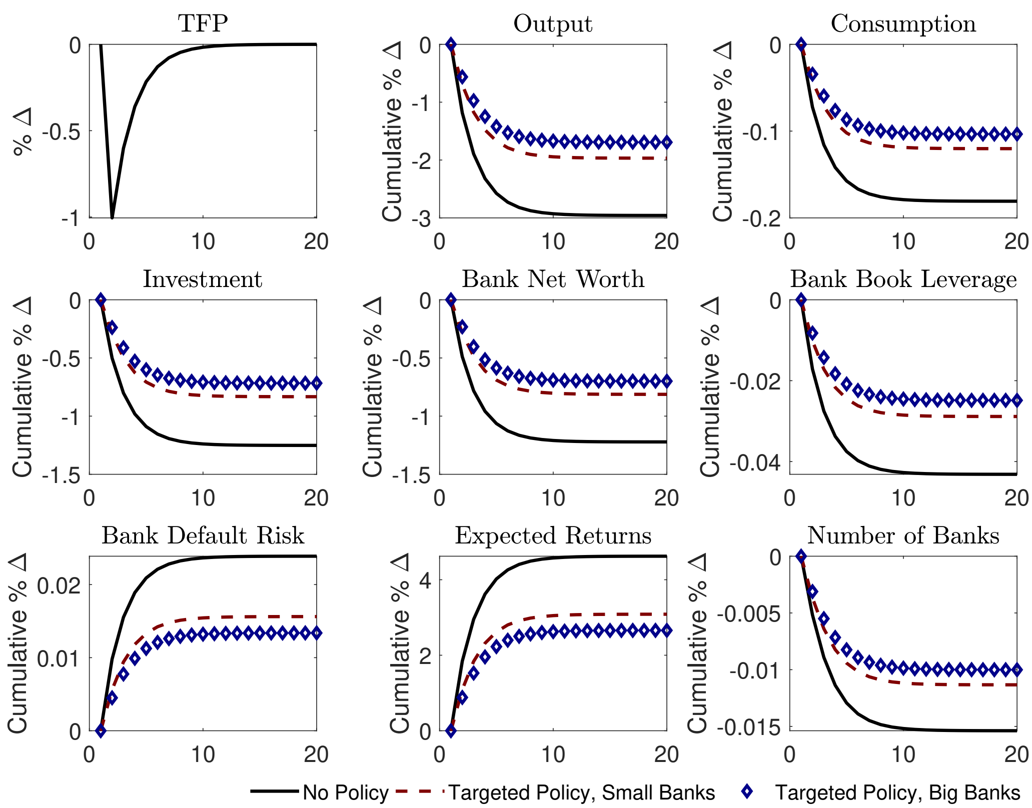
With $\bar{\Omega}$ and $\bar{\mu}$ fixed at the steady-state values. That is, the total direct effect comprises the aggregated partial-equilibrium response of credit supply to the exogenous disturbance alone without updating aggregate general equilibrium variables and the banking distribution.

The indirect effect from Equation 1.36 includes four distinct channels of transmission. First, aggregate productivity impacts demand for investment. Because firms require external financing 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 τ_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.

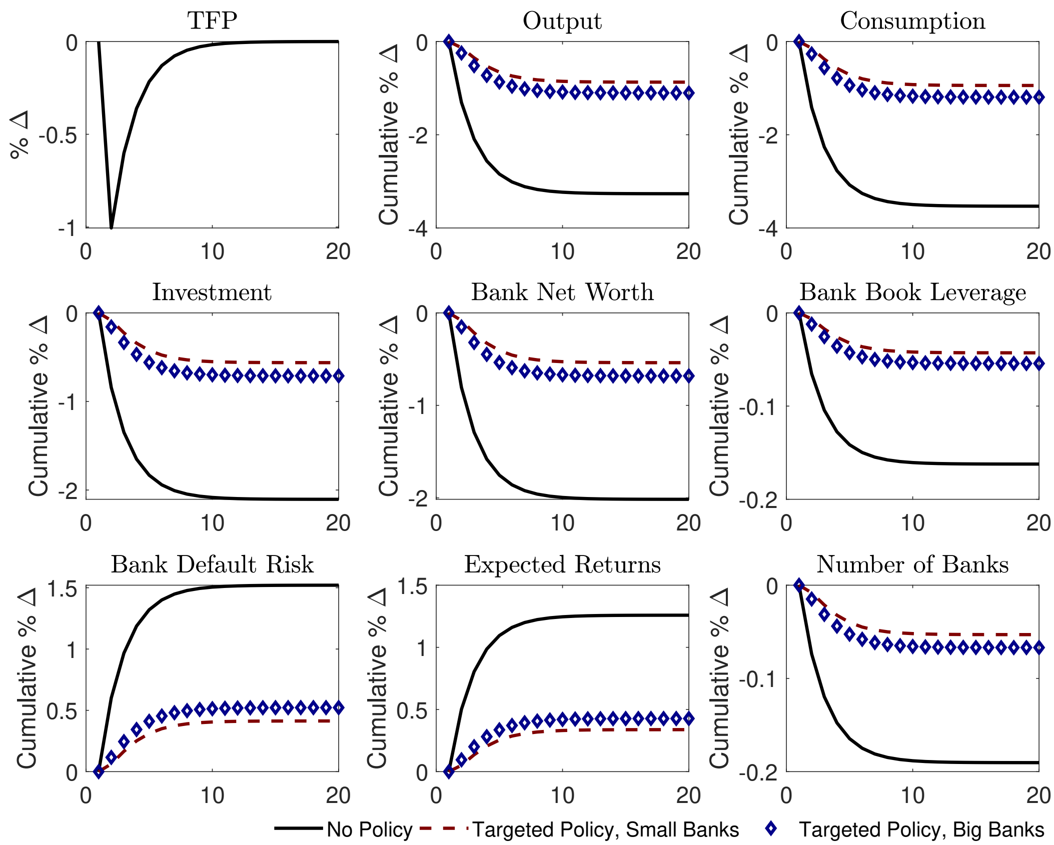
Figure 1.24: **Crisis Experiment: Targeted Equity Injections**



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 1.24 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 1.25: **Crisis Experiment: Liquidity Facility**



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

Figure 1.24 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 1.25 plots the same numerical experiment but now with targeted liquidity facilities. These policies reduce the fraction of divertible assets λ 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 λ 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).

1.10 Additional Model Details and Derivations

1.10.1 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 1.26. We analyze the optimal choice of bank market leverage $\frac{\mathrm{pk}}{\mathrm{n}}$ in three different situations. First, we start with the representative-bank case with complete markets (σ_{ξ}=0 and κ=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 1.26 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: ξ(j) and n(j): low-n(j), high-ξ(j) banks choose the highest leverage in the economy.

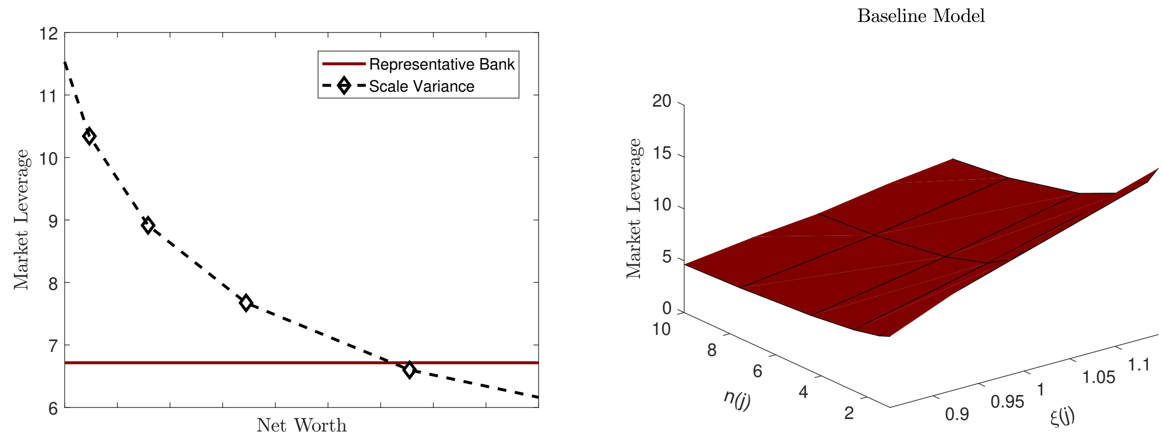
An important feature of this class of models with financial intermediaries is linearity with respect to net worth. This assumption normally allows the model to be aggregated explicitly. I can formalize the departure from homogeneity by formally proving that the value function of the bank in our model 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 $\{\mathrm{k(j)}, \mathrm{p(j)}, \mathrm{d(j)}\}$. 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* ξ(j)*, is*

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

where the marginal value of net worth is:

Figure 1.26: **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.

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

and the multiplier on the moral hazard leverage constraint is

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

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

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

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

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

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

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.[28] □

1.10.2 Bank Markups and Marginal Costs

In this section we provide a proof for Proposition 4. 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

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

[28]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 $Z := \left(\int_0^1 \Upsilon'\left(\frac{k(j)}{K}\right)\frac{k(j)}{K}dj\right)^{-1}$. The first order condition is

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

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:

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

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:

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

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).

□

1.11 Microfoundations and Extensions

1.11.1 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 ϵ_{ij} which is borrower-creditor specific.

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

$$Q_i\left(\bar{\epsilon}; k_i\right) = \underbrace{\log k_i + Y - p_i k_i}_{\text{Homogenous across j}} - \underbrace{\sum_{k=1}^{M}\left(\bar{\epsilon}^k - \epsilon_i^k\right)^2}_{\text{Heterogeneous across j}} \qquad i=1\ldots H \tag{1.38}$$

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:

$$\bar{E}^j = \frac{\log \frac{p(j)}{p_n}}{4} \tag{1.39}$$

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:

$$k_n = \frac{1}{p_n}\int_{-\infty}^{\bar{\epsilon}_1}\ldots\int_{-\infty}^{\bar{\epsilon}_{n-1}} f\left(\bar{E}^j\right) d\bar{\epsilon} \tag{1.40}$$

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

$$\frac{\partial^{H-1} k_n}{\partial p_1 \ldots \partial p_{H-1}} = \frac{1}{p_1 \cdots p_n} 4^{1-H} f\left(\bar{E}^j\right) \tag{1.41}$$

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:

$$f(\bar{\epsilon}) = H \frac{4}{\bar{\theta}}^{H-1} (H-1)! \frac{\prod_{i=1}^{H-1} \exp\left(-4/\bar{\theta}\epsilon^i\right)}{\left[1 + \Sigma(j)^{H-1} \exp\left(-4/\bar{\theta}\epsilon^j\right)\right]^H} \tag{1.42}$$

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

$$\frac{\partial^{H-1} k_n}{\partial p_1 \ldots \partial p_{H-1}} = H \bar{\theta}^{1-H} (H-1)! \frac{\Pi(j)^H p(j)^{-1/\bar{\theta}-1}}{\left[\Sigma(j)^H p(j)^{-1/\bar{\theta}}\right]^H} \tag{1.43}$$

Integrating gives us optimal credit demand

$$k_n = H p_n^{-1/\bar{\theta}-1} \left[\sum(j)^H p(j)^{-1/\bar{\theta}}\right]^{-1} \tag{1.44}$$

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:

$$k(j) = \left(\frac{p(j)}{P}\right)^{-\theta} K \tag{1.45}$$

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 θ. The relationship is inversed, so a higher $\bar{\theta}$ is associated with a lower elasticity of credit supply, i.e. greater credit market power. In the limit, if $\bar{\theta} \to \infty$ 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.

1.11.2 Portfolio Returns

In this section we explain how our formulation of total portfolio returns (Equation 3.9) 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:

$$\mathrm{R}^{\mathrm{b}} = \sum_{\mathrm{j}}^{\mathrm{N}} \frac{1}{\mathrm{N}} \mathrm{R}^{\mathrm{j}} + \kappa \mathrm{R}^{\mathrm{c}} - \kappa \mathrm{R}^{-\mathrm{c}} \tag{1.46}$$

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

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

Which is the same formulation that we used in Equation 3.9, except that in the model R^{c} is $\xi(\mathrm{j})$ and follows an autoregressive process. Now, total return across all banks can be written as:

$$\mathrm{R}^{\mathrm{total}} = \sum_{\mathrm{b}}^{\mathrm{N}} \frac{1}{\mathrm{N}} \kappa \mathrm{R}^{\mathrm{c}} + (1-\kappa) \sum_{\mathrm{b}}^{\mathrm{N}} \mathrm{R}^{\mathrm{k}} = \mathrm{R}^{\mathrm{k}}$$

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 $\xi(\mathrm{j})$.

1.11.3 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:

$$\min_{K_{at}, K_{bt}} P_t K_{at} + K_{bt} \quad \text{s.t.} \quad K_{at}^{\chi} K_{bt}^{1-\chi} = K_t \tag{1.47}$$

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 χ. The second stage of the problem is determination of the demand for individual varieties $k_t(j)$ within the K_{at} sector.

The parameter χ 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 χ could be used to simulate advancements in financial innovation and/or the rise of complexity in the credit market.

1.12 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.

1.12.1 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 $\{\upsilon'(j), \bar{R}'(j)\}$.

7. Solve the household's problem using time iteration. Get new Λ'.

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.[29]

1.12.2 Constrained Efficient Equilibrium

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

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

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.

[29] 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 Λ'.

9. Compare $\{\alpha^{k}, \alpha^{p}\}$ with $\{\alpha^{k'}, \alpha^{p'}\}$. And compare $\{\bar{R}(j)\}$ with $\{\bar{R}'(j)\}$. If the maximal errors across all grid points are within the tolerance level, 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.

1.13 Data Description

Empirical data used for model validation is obtained from the U.S. Call Reports. Table 1.5 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 1.5: **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.

Chapter 2

Granular Credit Risk

What is the impact of granular credit risk on banks and on the economy? We provide the first causal identification of single-name counterparty exposure risk in bank portfolios by applying a new empirical approach on an administrative matched bank-firm dataset from Norway. Exploiting the fat tail properties of the loan share distribution we use a Gabaix and Koijen (2020a,b) granular instrumental variable strategy to show that idiosyncratic borrower risk survives aggregation in banks portfolios. We also find that this granular credit risk spills over from affected banks to firms, decreases investment, and increases the probability of default of non-granular borrowers, thereby sizably affecting the macroeconomy.

2.1 Introduction

What is the impact of idiosyncratic borrower risk on banks and the economy? It has been understood for years that if individual loans are small enough relative to the overall size of the portfolio then credit risk pooling should achieve perfect insurability against idiosyncratic shocks (Diamond, 1984). But what if some loans are large? What if the distribution of loan sizes is fat-tailed: can the performance of a single large loan directly affect portfolio-level outcomes and lending? A rapidly growing literature, originating from the seminal contribution by Gabaix (2011), has emphasized the micro - or "granular" - origins of macroeconomic outcomes in a variety of theoretical and applied contexts. According to the granular hypothesis, shocks to large, non-atomistic agents generate non-diversifiable "grains" of economic and financial activity, which can directly affect aggregate fluctuations and, via general equilibrium effects, all other agents.

This working paper should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. We thank our discussants José-Luis Peydró and Daniel Paravisini as well as Christoph Basten, Svetlana Bryzgalova, Andreas Fagereng, Julian di Giovanni, Francisco Gomes, Refet Gürkaynak, Victoria Ivashina, Joseba Martinez, Atif Mian, Steven Ongena, Elias Papaioannou, Anna Pavlova, Kasper Roszbach, Stephen Schaefer, Vania Stavrakeva, Kjetil Storesletten, Paolo Surico, Anette Vissing-Jorgensen, Gianluca Violante and seminar participants at the CEPR/ERC/LBS Conference on Granularity and Applications, AFA 2021, EEA 2020, SNDE 2020, WEAI 2020, LBS, Norges Bank, Nordea Bank, Oslo Macro Group, Statistics Norway and the University of Zurich for valuable comments and suggestions. We thank Mikkel Riiser for excellent research assistance. Rey is very grateful to the ERC (Advanced Grant 695722). All errors are our own.

Curiously, there are few empirical applications of the granular hypothesis to banking. This is puzzling because in practice the hypothesis maps directly into the "large exposure regulation" of the Basel Committee on Banking Supervision (BCBS). The BCBS has been regulating bank credit concentration risk for decades, formally at least since the Basel I Accords. The *Core Principles for Effective Banking Supervision* emphasize that local country laws should "set prudent limits on large exposures to a single borrower" (BIS, 2013). In practice however, the Principles admit that "material differences in scope of application, the value of large exposure limits, methods for calculating exposure values, and more lenient treatments for certain types of exposures exist". As a result, the document concludes, "although a concentration risk adjustment could be made to mitigate these risks, these adjustments are neither harmonised across jurisdictions, nor designed to control traumatic losses from a single counter-party default".

This paper is the first to provide causal empirical evidence on the importance and implications of "single-name" credit concentration risk[1]. We develop a new empirical approach and apply it to a novel administrative firm-bank matched dataset from Norway[2]. We merge our loan-level administrative database with firm and bank balance sheet data. We cover every single bank loan made to limited liability companies (LLC) in Norway over the 2003-2015 period[3]. This data-rich environment enables us to study the transmission mechanism and heterogeneous treatment effects at many levels of aggregation.

Our empirical strategy consists of five steps. First, we establish that the distribution of loan shares in our dataset is fat-tailed. Our estimate of the Pareto power implies that 80% of all credit is concentrated in 20% of the loans. Interestingly, we provide therefore another example of the famous "80-20" Pareto principle that occurs in a variety of settings in economics as well as more generally in social and physical sciences (Gabaix, 2009).

Second, we construct a measure of idiosyncratic borrower risk. We use data on firm balance sheets and income statements to estimate idiosyncratic value-added shocks for the universe of all LLC firm × years in Norway over 2003-2015. We extract non-systematic variation in firm value-added by controlling for a variety of balance sheet items like firm size and costs as well as firm, industry, year, and geographical fixed effects. Our approach follows very closely a large literature in labor economics and macroeconomics (Guiso et al., 2005; Hsieh and Klenow, 2009; Fagereng et al., 2018)[4].

[1] We follow the BCBS vocabulary where "single-name"refers to the level of an individual borrower or counterparty. This is in contrast, for example, to how BCBS defines and treats sectoral or geographical exposures where the unit of analysis is either a whole industry or region.

[2] Throughout the paper we focus on corporate clients and loans. Our empirical approach however, is general and flexible enough to be applied to other borrower types such as households, state institutions, or other intermediaries.

[3] LLC is by far the most commonly used organizational structure in Norway. For most years, our firm data accounts for more than 90% of total employment in the private sector.

[4] We perform a variety of validation and robustness checks to discipline our measure. First, we run a series of factor

Third, we establish the pass-through from these idiosyncratic firm shocks to loan-level returns. We investigate how such shocks affect returns on loans within the same bank, industry, county, year, and loan type. Importantly, our specification controls for any confounding bank-side supply factors, potentially specific to a given industry, county, or contractual type[5]. We find that idiosyncratic firm shocks have a strong effect on loan returns. In our preferred specification with a full set of controls and fixed effects, a one standard deviation negative firm shock causes annual loan-level returns to fall by 36 basis points. We explore numerous dimensions of heterogeneity, including firm characteristics, geographical location, ownership, and sector.

Fourth, we look at the impact of idiosyncratic borrower shocks on banks portfolio-level outcomes. This is a critical step in our analysis. Once aggregated to the level of a bank, we potentially lose the positive properties of loan-level analysis: the loan share-weighted firm shock series could be potentially contaminated by bank $\times$ year confounding factors which we no longer have the power to deal with. For this stage, we adopt the "Granular Instrumental Variable" (GIV) approach, newly developed in a series of papers by Gabaix and Koijen (2020a,b). Intuitively, the GIV extracts the variation in the share-weighted aggregated firm shock series that can be attributed to "granular" borrowers. Specifically, the instrument in its simplest form is the difference between size-weighted and unweighted aggregated firm shocks. The GIV thus purges away any bank $\times$ year factor almost by construction. Conditional on the distribution of credit shares being fat-tailed, idiosyncratic shocks to large borrowers allow us to achieve identification. Our various parameteric and non-parameteric specifications allow for a flexible number of bank factors and, importantly, for loadings on bank factors to be either homogenous or *heterogeneous* across firms within any bank's portfolio.

One important result of our paper is that idiosyncratic firm shocks, instrumented by the GIV, have a large and significant effect on portfolio-level return on loans (RoA). A one-standard-deviation granular credit shock causes portfolio RoA to move by 11.6 basis points on average. Given that in the estimation sample the standard deviation of RoA is 1.35, our estimate can explain 8.6% of the total dispersion of bank returns. We also find that the relationship is strongly concave, driven mainly by negative shocks. In particular, if we condition on positive share-weighted shocks, the estimated coefficient becomes a noisy zero. In contrast, when conditioning on negative share-weighted shocks, the estimate jumps to as high as 19.4 basis points, which is 15% of the sample standard deviation of

---

analysis exercises whereby we explicitly extract parametric and nonparametric common factors. Second, we establish that shocks are not cross-sectionally correlated or persistent across time. Third, we run several placebo permutation tests. Fourth, we show that these shocks only have contemporaneous and lead effects on loan and bank outcomes, i.e. that there are no "pre-trends". Finally, we validate our measure with a narrative-based approach by matching realizations in the bottom percentile of the shock distribution to actual news stories from Norwegian media.

[5]Conceptually, this step can be viewed as a "reverse Khwaja and Mian (2008)" approach. In Khwaja and Mian (2008), authors trace out the impact of bank supply shocks onto firms that borrow from the same bank. This way, they are able to control for any confounding firm-side factors. Our strategy is to compare loan outcomes within the same bank in order to control for supply-side factors. Our approach is very "granular", since we zoom in on firms within the same bank, *and* industry and county.

RoA - an increase of 74% over the average estimate[6]. We investigate heterogeneity at the bank level and find that the pass-through of granular credit shocks is stronger for banks with high portfolio risk weights, low assets, high loan portfolio concentration, and high profitability. Furthermore, conditional on the sample of banks with high risk weights, the pass-through is especially strong if banks are small, profitable, and hold concentrated portfolios. We also find that the *number* of loans in credit portfolios does not affect the transmission mechanism, indicating that granular credit risk is not merely a "small-N" problem.

Fifth, having established that shocks to granular borrowers have a direct effect on portfolio-level returns, we ask whether banks pass on these shocks to the real economy. In other words, are there macroeconomic spillovers from granular credit risk? We start by examining credit supply effects, by comparing bank loan quantity and rate changes in response to granular credit shocks. We restrict the sample to firms with multiple bank relationships, and ask if banks that experience bad granular credit outcomes reduce credit supply or increase prices. The within-firm analysis allows us to control for demand side effects using time-varying firm fixed effects, thus isolating the supply side. We find strong evidence, both in terms of quantity and price effects, that banks pass on granular credit shocks to their *non-granular* clients, i.e. firms with a loan share that is less than a certain threshold (such as the median) in the pooled distribution of all credit shares[7]. We show that a one-standard-deviation bank-level negative granular credit shock reduces loan supply and increases interest rates by as much as 71.7 and 63.4 basis points, respectively. This identifies a leftward shift of the credit supply curve: quantities fall while prices rise. There are "granular credit risk spillovers": firms are affected by granular shocks of others via their bank.

We then ask whether affected non-granular firms experience negative real economic outcomes. We find that affected non-granular firms cut investment. Moreover, these firms experience elevated bankruptcy rates for up to 3 years after the initial shock. A one-standard deviation negative granular credit supply shock increases the likelihood of bankruptcy by roughly 32-60 basis points for all firms, and 68-101 basis points for non-granular borrowers. Granular credit risk has therefore sizable implications for the aggregate economy.

An important question is whether banks hedge granular risk with alternative sources of income.

---

[6]The concave relationship is reassuring to us for the simple reason that it reflects the basic payoff structure of the debt contract. While there is no upside for the lender from borrowers experiencing positive value-added shocks, the downside is capped only by the principal of the loan, not counting default-related costs, be they pecuniary or not. Apart from the intuitive economic interpretation, we also view our finding of strong asymmetric effects as an important sign of validation that our measure of idiosyncratic shocks is indeed economically informative.

[7]The tendency to pass along adverse economic shocks to their clients, especially small firms, is not uncommon for banks. In their classic paper, Peek and Rosengren (2000) find that the 1990s Japanese banking crisis had a negative effect on commercial real estate activity in the U.S. through the network of banks exposed to both markets. Lin and Paravisini (2012) trace out the pass-through of the collapse of WorldCom on firms that shared the same lender. In a recent paper, Greenwald et al. (2020) show that banks that experience larger credit line drawdowns restrict lending to firms that borrow through term loans - a negative spillover effect on smaller borrowers.

For example, in states of the world where credit income is low derivative income could be high. We collect detailed bank-level data on non-interest income and find that none of the measures we have correlate with GIV-instrumented firm shocks. We see no correlation between our shock measures and fees income, equity and bond appreciations, dividend income, or derivatives income. Another issue is that banks could potentially pre-insure against granular borrower shocks by charging higher markups for risky clients. But when extracting value-added shocks we control for firm size, liquidity, credit ratings, leverage and time-invariant factors. From the bank's perspective, unless markups are stochastic for some very unique reason, any firm shock comes therefore as a surprise relative to the firm's average performance and its creditworthiness, both of which would in turn be very highly correlated with the loan contract's price.

Finally, we supplement our empirical analysis by providing a theoretical motivation. We introduce parsimoniously bank credit into the canonical framework of Gabaix (2011). We model firm borrowing needs as a power function of firm size, which in turn is drawn from a power law density. Under this assumption, the distribution of bank loans (or, equivalently, firm borrowing) follows the Singh-Maddala (SM) family of densities (Singh and Maddala, 1976). The SM distribution has been used extensively to model wealth and income inequality. Our main contribution here is the derivation of sufficient statistics-based parameter restrictions under which the bank loan distribution also has a fat tail. If that is the case, then shocks to large borrowers may survive aggregation and impact bank-level portfolio outcomes. Using our dataset, we provide maximum likelihood estimates of the sufficient statistics and confirm that all the restrictions are satisfied on average.

Literature Review Our paper relates to several literatures. First, it builds on the rapidly growing literature on the "granular hypothesis" and its applications. Some of the more salient contributions across fields range from papers on business cycles (Carvalho and Gabaix, 2013), to trade (Gaubert and Itskhoki, 2018), international finance (di Giovanni et al., 2018), asset management (Choi et al., 2017), and banking (Amiti and Weinstein, 2018; Bremus et al., 2018). The latter strand focuses on how idiosyncratic bank *supply* shocks can have aggregate real implications whereas we focus on the transmission of shocks to (large) borrowers onto banks and the real economy. Our contribution is to show the existence of important granular credit risk spillovers on the economy.

Second, we relate to the literature studying the trade-off between credit concentration and diversification. On one hand, diversification enhances credit monitoring and information provision capacity (Diamond, 1984; Boyd and Prescott, 1986). On the other hand, some empirical studies found a positive correlation between portfolio concentration, returns, and monitoring efficiency (Acharya et al., 2006). Beck et al. (2017) have shown that bank specialization and concentration potentially have positive implications for systemic financial stability. Our paper contributes to this debate in at least two ways. We argue that as long as the distribution of credit shares features a fat tail, banks remain exposed to idiosyncratic shocks to their (granular) borrowers. Everything else equal, this is detrimental for financial stability. Because we find that banks pass on granular credit shocks to the real economy, credit concentration induces negative economic outcomes on average, *ceteris paribus*. But a normative interpretation of our results depends on the precise theories generating loan concentrations in the first place, an issue we discuss further in Section 2.6.

There is an emerging new literature on credit concentration that, like us, takes advantage of detailed microeconomic data. In a recent study, Agarwal et al. (2020) find that Mexican banks that specialized in energy lending around the 2014 collapse of energy prices amplified the sectoral shock to the rest of the real economy. Paravisini et al. (2020) find that persistent bank market-specific specialization can explain a significantly larger fraction of within-firm variation in credit than actual bank supply shocks. Goetz et al. (2016) show that geographic diversification by banks has no impact on average loan quality and is associated with a reduction of exposure to local idiosyncratic risks. Finally, Huremovic et al. (2020) and Dewachter et al. (2020) study the role of production networks in Spain and Belgium, respectively, for the propagation of bank shocks[8]. Our paper differs from this literature because we work explicitly with *single-name* concentration risk, while most of the literature deals with either sectoral or geographical specialization. In addition, we emphasize both empirically and theoretically the importance of granularity of the loan share distribution for the pass-through of idiosyncratic shocks to the aggregate bank portfolio. Thus our paper provides an empirical basis for the work of Mendicino et al. (2019) who show in a quantitative model that if banks are not perfectly diversified, the interaction between borrowers' and banks' solvency has important effects on the probability and severity of crises.

The remainder of the paper is structured as follows. Section 2.2 provides a description of our data. Section 2.3 describes the different stages of our empirical approach. Section 2.4 reports the main empirical results. Section 2.5 explores heterogeneity at different levels of aggregation and reports results from various robustness checks. Section 2.6 discusses the assumptions we make in our empirical approach and the implications of our findings. Section 2.7 lays out our theoretical motivation. Finally, section 2.8 concludes.

2.2 Data

Our empirical investigation is based on a unique dataset assembled from three major sources: administrative data from the Norwegian Tax Authority, credit rating agency data from Bisnode and supervisory data from ORBOF. They were merged using the unique identifiers for banks and firms. The Norwegian Tax Authority data is a high-quality matched firm-bank administrative register.

[8]We do not have information on production linkages in our data so we cannot explore this potentially complementary channel of propagation.

The unit of observation in this database is an individual loan and the frequency is annual. For every loan, we observe the firm-bank identifiers as well as the flow of interest paid during the year and the end-of-year stock of debt. Because the data is collected and maintained by the tax authority as a basis for corporate taxation, the variables are essentially measurement error-free.[9] The data set covers all limited liability companies for the time period of 2003-2015, which accounts for roughly 90% of private sector employment for most years. We aggregate all loans into a single annual firm-bank "relationship" unit. The terms loan and relationship are used interchangeably, and refer to the sum of loans and interests paid across all individual loans between a bank and a firm.

A key measure in our analysis is the return on a loan, or a credit relationship (RoL). This is not directly observed, and hence we impute it. Specifically, we observe interest collected throughout year t (R_t) and the end-of-year stock of outstanding debt (D_t). We then define the RoL in year t as $R_t/(0.5D_{t-1} + 0.5D_t)$, which is equivalent to interest received relative to the average of debt outstanding at the beginning and end of the calendar year.

We merge the loan-level data with detailed information on Norwegian firms and banks. Our firm data comes from the credit rating agency Bisnode. In addition to information about the firms' credit rating scores and firm characteristics such as age, location and industry, the data set includes annual balance sheet and income statement items on all Norwegian firms for 1999-2019. The bank data is from a supervisory registry (ORBOF) and includes annual balance sheet and income statement information covering all Norwegian banks over 1987-2019. The data set also provides us with confidential information on non-interest income, including income from derivatives, equity and bond investment, dividends, and loan fees.

We perform several cleaning and truncation steps on the raw data. First, we drop observations that are clearly erroneous, such as cases of liquidity ratios being greater than 1. Second, following Foster et al. (2008) we truncate the distribution of cost-to-total-cost ratios for each cost type at the 10% and 90% in each industry and year. Cost types include wage bill, energy, material and other costs. This is important as firms could dump all their operational costs to a particular fiscal year in order to receive tax advantages, and what we would thus pick up are in fact endogenous outcomes rather than unanticipated performance shocks. Third, we truncate the extracted firm shock distribution at the 1% and 99% levels. All our main results at the loan and bank levels are quantitatively robust to alternative cleaning rules. Table 2.1 provides summary statistics for some of the key variables used in our analysis.

[9]Provision of false tax information carries substantial legal, financial and reputational penalties. Additionally, the information about outstanding debt and interest paid is reported to the tax authority by the banks, and not the firms themselves.

Table 2.1: **Descriptive Statistics**

Variable	Observations	Mean	Std. Dev	Min	Max
		Loans			
Interest Received	333289	196645.31	1620919.78	1.00	2.67e+08
Loan Amount Outstanding	333289	4035259.25	43884811.59	1.00	7.00e+09
Return on Loan	333289	9.01	8.92	0.00	100.00
		Firms			
Sales (1000 NOK)	277707	26532.69	217768.69	0.00	33761000.00
Total Assets (1000 NOK)	277707	42361.08	1052017.18	2.00	1.20e+08
Wage Costs (1000 NOK)	277707	6827.88	65057.01	1.00	7098000.00
Material Costs (1000 NOK)	277707	11643.95	103640.10	0.00	15313000.00
Equity / Assets Ratio	277707	0.27	0.18	0.00	1.00
Liquidity Ratio	277707	0.16	0.17	0.00	1.00
Employees	277707	15.81	156.66	0.00	20781.00
Firm Age	277707	12.94	11.81	0.00	159.00
		Banks			
Return on Loans	1380	6.40	1.46	0.06	14.39
Total Assets	1377	21130037.71	1.35e+08	92384.00	1.96e+09
Total Equity	1377	1491611.98	8512785.73	16139.00	1.51e+08
Assets / Equity Ratio	1377	10.90	3.20	1.32	41.48
Cash Balances / Assets	1377	0.03	0.03	0.00	0.33
Number of Loans	1380	220.88	854.18	1.00	8940.00
Loan Herfindahl Index	1380	0.10	0.12	0.00	1.00
Share of 10% Largest Loans	1380	0.54	0.13	0.20	1.00
Share of 5 Largest Loans	1380	0.51	0.20	0.07	1.00
Deposits to Assets Ratio	1377	0.66	0.12	0.01	0.91
Financial Assets Ratio	1321	0.08	0.06	0.00	0.48
		Estimated Idiosyncratic Shocks			
Firm-level	277707	0.02	0.27	-1.42	1.15
Bank-level (size-weighted)	1380	-0.02	0.11	-0.78	0.69
Granular IV	1380	-0.02	0.09	-0.76	0.46

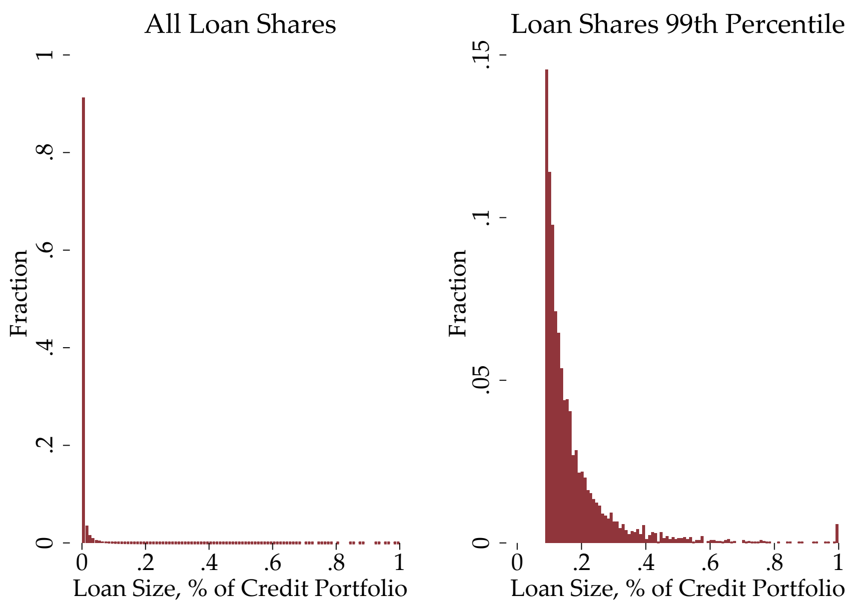
Notes: This table shows summary statistics of key loan, firm, and bank characteristics. All stock and earnings values are in thousands of Norwegian Kronas (NOK). 1 US Dollar = 8.30 NOK as of June 4, 2021. Firm shocks in panel 2 are estimated according to specification 2.1. Loan data is from the Norwegian Tax Authority. Firm data is from the credit rating agency Bisnode. Bank data is from the financial supervisory database ORBOF. Sample includes all bank loans to limited liability companies in Norway over 2003-2015.

2.3 Empirical strategy

2.3.1 Granularity of the Distribution of Loan Shares

We begin by establishing that the distribution of loans shares in our dataset is fat-tailed. In Figure 2.1 we plot the histogram of all loan shares, pooled across all banks and years over 2003-2015. Eyeballing the distribution is enough to notice its extreme skewness. More formally, we fit the Pareto I density to the data and estimate a Pareto rate of 1.16. Any estimate below 2 implies that idiosyncratic shocks to large loans potentially survive risk pooling and cause portfolio-level disturbances. This follows directly from the proofs in Gabaix (2011). Interestingly, our estimate of the Pareto power implies that 80% of all credit is concentrated in 20% of the loans. Thus the loan share distribution provides yet another example of the famous "80-20" Pareto principle that occurs in a variety of settings in economics as well as in many social and physical sciences applications (Gabaix, 2009). In section 2.7, we introduce a parsimonious model of bank credit into the canonical framework of Gabaix (2011). In our model, the fat tail of the firm size distribution feeds directly into the fat tail of the loan share distribution under certain parameter restrictions. We estimate the main parameters of the model using our data and confirm that those restrictions are on average satisfied.

Figure 2.1: **Distribution of Bank Loan Shares**



Notes: This graph presents the distribution of bank loan shares. The left picture plots the full distribution. The right picture zooms in on the 99th percentile of the shares. The share of each loan is computed as the ratio of a singular loan's amount to total corporate loans of a given bank in a given year. The figures plot the pooled shares for all banks and years. The Pareto rate of the 99th percentile is 1.16.

2.3.2 Estimates of Idiosyncratic Firm Shocks

The next step of our empirical approach consists of extracting idiosyncratic firm shocks, measured as unexplained idiosyncratic variation in firm value-added. Our approach follows closely a large number of studies in labor and macro economics that extract idiosyncratic sales or performance shocks. (Foster et al., 2008; Hsieh and Klenow, 2009; di Giovanni et al., 2014; Foster et al., 2017; Fagereng et al., 2018).[10] To extract unexplained variation in firm value-added, we regress the log of firm value-added on a set of time-varying firm-level controls that includes measures of input usage and firm riskiness. Importantly, since our focus is on idiosyncratic variation, we remove common (across firms) components by controlling for the interaction of time, industry and county fixed effects. Finally, across-firm variation attributed to time-invariant firm characteristics is absorbed by firm fixed effects.

Formally, for a firm j, operating in an industry s from a county z in year t, we estimate the

[10]Using idiosyncratic shocks as "instruments" for estimating microeconomic or macroeconomic elasticities is increasingly common in applied microeconomics and finance (see Leary and Roberts (2014), Amiti et al. (2019) and Gabaix and Koijen (2020a)).

following regression:

$$\ln \mathrm{VA}_{j,t} = \alpha_{j,t,s(j),z(j)} + \beta_1 \ln \mathrm{K}_{j,t} + \beta_2 \ln \mathrm{W}_{j,t} + \lambda' \mathrm{X}_{j,t} + \epsilon_{j,t} \tag{2.1}$$

where VA stands for firm value-added[11], K represents book capital, W the wage bill, and X are other controls including leverage, liquidity, credit rating, and a quadratic polynomial in age. The term $\alpha_{(\cdot)}$ captures a combination of fixed effects at the firm and industry × year × county levels. Here, K and W are proxies for capital and labor inputs, while X are various measures of firm riskiness. These factors should capture the banks' information set well.[12] In addition to the specification in (2.1) we also consider a less conservative specification, which only includes fixed effects but not any of the other controls, in the spirit of di Giovanni et al. (2014).

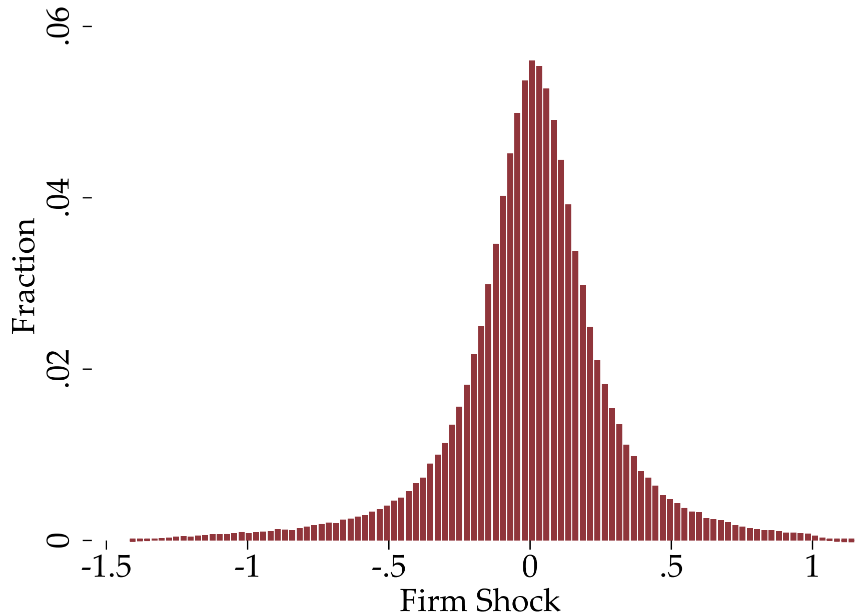
The object of interest is the residual from this regression, $\epsilon_{j,t}$, which is the main right-hand side variable for the rest of the paper. Essentially, what we are trying to capture are unforeseen changes in firm performance that banks, despite observing multiple layers of data, could not have anticipated. Examples of such events include a factory collapse, fraud and mismanagement, operational and logistical accidents, human error, etc. In Section 2.12 of the Online Appendix we provide a headlines and narrative-based explanation for some of the most negative shock realizations in our sample.

Figure 2.2 plots the distribution of our baseline shock measure $\epsilon_{j,t}$, pooled across all firms and years. It is noticeably left-skewed, with a larger mass in the left tail.

[11] Value added is measured as sales minus material, energy, and other costs.

[12] A potentially important factor that is missing from this specification is market prices. The share of publicly traded firms in our data is, however, very small. Moreover, credit rating arguably captures the same information that would be embedded in the stock price (albeit updated far less regularly).

Figure 2.2: **Distribution of Idiosyncratic Firm Shocks**



Notes: This graph plots the pooled distribution of idiosyncratic firm shocks estimated from equation (2.1).

Factor Analysis and Robustness Despite controlling for a variety of firm characteristics and fixed effects, there is still concern that our shocks $\epsilon_{j,t}$ may pick up some latent common components. In Section 2.9.1 of the Online Appendix, we generalize the reduced-form specification in (2.1) and formally extract parameteric and non-parameteric common factors from the residual $\epsilon_{j,t}$. All our results and insights at the loan and bank levels will remain unchanged. We also conduct a series of additional robustness tests in order to establish that our shocks are truly idiosyncratic. First, we confirm that $\epsilon_{j,t}$ are not correlated across firms or time. Second, we run several placebo regressions. We return to these robustness checks in Section 2.5.

2.3.3 Loan Outcomes

To identify the impact of idiosyncratic firms shocks on loan-level returns, we exploit the granular nature of our dataset. Individual bank-firm relationships enable us to control for time-varying bank supply factors, such as risk aversion or monitoring skills, by including bank $\times$ year fixed effects. Bank supply factors could confound our demand-side shocks.[13] We also control for interacted county $\times$ year $\times$ industry fixed effects. This specification implies that the impact of shocks is identified by comparing loan-level returns across firms in the same county, industry,

[13]Coimbra and Rey (2019), among others, show that heterogeneity in risk appetite among financial intermediaries is a determining factor for financial and business cycles. Our fixed effects specification takes care of this issue.

year, who are borrowing from the *same* bank. For some firm-bank relationships in our dataset we also observe the fraction of total loan volume that comes from credit lines. This allows us to also consider specifications which include a loan type fixed effect[14]. Formally, we estimate the following specification:

$$R_{i,j,t} = \alpha_{i,t,s(j),z(j),l(i,j)} + \beta\epsilon_{i,j,t} + \nu_{i,j,t} \tag{2.2}$$

where i is a bank that lends to firm j from county z, industry s, year t via loan type l. $R_{i,j,t}$ is the loan-level return and $\epsilon_{i,j,t}$ is the estimated idiosyncratic shock of firm j in bank i's portfolio. Because the main RHS variable is *estimated*, our standard errors are corrected for the estimated regressor bias via bootstrapping. Importantly, our specification features a wide range of fixed effects captured by the term $\alpha_{(\cdot)}$. Specifically, in our most conservative specification we include the full interaction of bank × year × firm industry × firm county × loan type fixed effects.

2.3.4 Granular Credit Risk: Bank outcomes

After investigating how idiosyncratic firm shocks affect loan returns, we then move up to the level of a bank portfolio. We aggregate realized idiosyncratic firm shocks to the bank level by weighing shocks with loan shares and refer to the resulting measure as "granular credit risk". Intuitively, granular credit risk captures shocks to banks' clients that eventually do not average out and instead impact portfolio-level outcomes.

To evaluate the bank-level impact, we proceed by analyzing the following relationship between bank-level returns on all corporate loans $R^b_{i,t}$ and firm shocks for bank i at year t:

$$R^b_{i,t} = \alpha_i + \alpha_t + \beta\bar{\epsilon}_{i,t} + \nu_{i,t} \tag{2.3}$$

where α_i and α_t denote bank and time fixed effects, $\bar{\epsilon}_{i,t} = \sum_{j\in J_i} s_{i,j,t}\epsilon_{i,j,t}$ are bank-level aggregates of firm shocks that are weighted by loan shares $s_{i,j,t}$, and $\nu_{i,t}$ the error term. The portfolio loan return $R^b_{i,t}$ is computed as the loan-share weighted average of loan-level returns.

There is one key identification challenge associated with the naive specification above. Our loan-level analysis exploited within-bank-year variation to control for confounding credit supply shocks. This is no longer possible when we turn our focus to outcomes at the bank level. Consider a generic time t relationship between bank outcome $y_{i,t}$, unobserved bank-side factor $\eta_{i,t}$, and demand-side idiosyncratic firm disturbance $\epsilon_{i,j,t}$:

$$y_{i,t} = \beta \sum_j s_{i,j,t}\epsilon_{i,j,t} + \varphi_i\eta_{i,t} \tag{2.4}$$

[14]A firm-bank relationship is classified as a credit line loan in year t if more than 50 percent of total credit in the relationship comes from credit lines.

where $s_{i,j,t}$ is the normalized exposure of bank i to firm j ($\sum_j s_{i,j,t} = 1$). Estimation of Equation 2.4 may potentially result in a biased estimate of β if $\eta_{i,t}$ and $\epsilon_{i,j,t}$ are correlated.

In order to alleviate this concern, we adopt a newly proposed "granular instrumental variable" (Gabaix and Koijen, 2020a) approach. Specifically, we assume that the demand-side disturbance $\epsilon_{i,j,t}$ can be written as

$$\epsilon_{i,j,t} = \delta_i \eta_{i,t} + u_{i,j,t} \tag{2.5}$$

where δ_i is the factor loading.

The granular instrumental variable ("GIV") is defined as the time-varying difference between exposure-weighted and equally-weighted firm shocks, each aggregated to the bank level. This way, the bank-time supply-side factor $\eta_{i,t}$, which is potentially correlated with firm disturbances, is purged out. The GIV is formally constructed the following way:

$$GIV_{i,t} = \sum_j s_{i,j,t}\epsilon_{i,j,t} - \sum_j \frac{1}{N_i}\epsilon_{i,j,t} = \sum_j s_{i,j,t}u_{i,j,t} - \sum_j \frac{1}{N_i}u_{i,j,t} \tag{2.6}$$

where N_i denotes the number of firm exposures of a given bank i. We now replace the naive approach in equation (2.3) with the following specification:

$$R^b_{i,t} = \alpha_i + \alpha_t + \beta \hat{u}_{i,t} + \nu_{it} \tag{2.7}$$

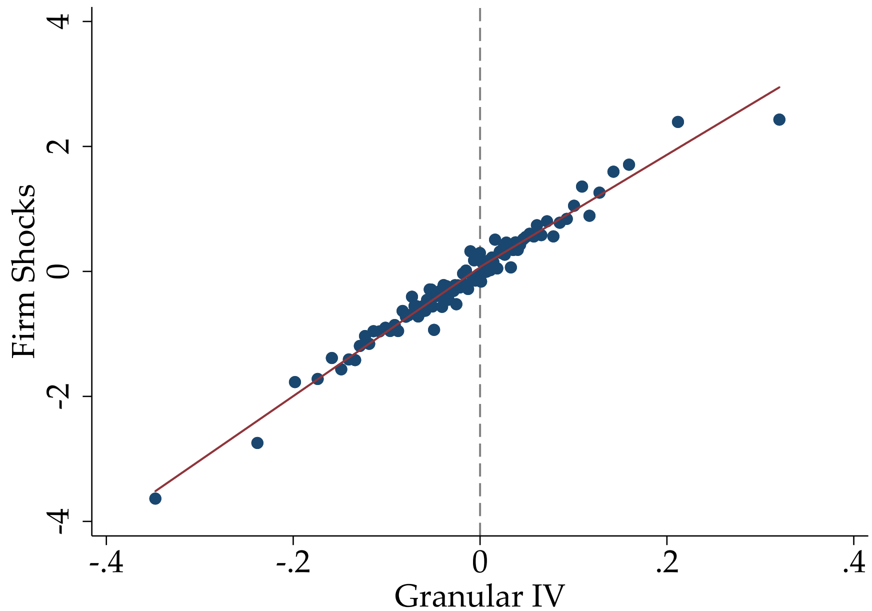
where $\hat{u}_{i,t}$ is the fitted value from the "first-stage" regression of the endogenous covariate $\bar{\epsilon}_{i,t}$ on the granular instrument $GIV_{i,t}$. This way, all variation in $\bar{\epsilon}_{i,t}$ is driven by fluctuations originating from the "granular borrowers", i.e. those with a large credit share. Naturally, if there are no granular borrowers, this approach does not work as there is no variation in the instrument. But as we have seen from Figure 2.1, the distribution of loan shares is very skewed. The main identifying assumption of this empirical approach is the following condition:

$$\sum_j^N \mathbb{E}\left[s_{i,j,t}u_{i,j,t}\nu_{i,t}\right] = 0 \tag{2.8}$$

for all i and t.

In words, identification is achieved if firm shocks $s_{i,j,t}u_{i,j,t}$ are not driven by the error term in bank-level regressions. This is the "exclusion" assumption. The main concern is that loan shares could be endogenous, i.e. correlated with firm shocks. This is not a problem for us for several reasons. First, credit demand in absolute terms correlates with firm size. Given how we extract idiosyncratic value-added shocks (i.e. controlling for size), our shock series is mechanically orthogonal to firm size. Similar logic applies to other firm factors such as leverage, liquidity,

Figure 2.3: **First Stage - Firm Shocks and the Granular IV**



Notes: This figure plots the relationship between the endogenous covariate $\bar{\epsilon}_{i,t}$ and the instrument, $GIV_{i,t}$. On the vertical axis we have the idiosyncratic firm shock which is loan size-weighted and aggregated to the level of a bank. Idiosyncratic firm shocks are extracted from specification 2.1. The granular instrument (horizontal axis) is constructed based on equation (2.6). Correlation between the two variables is 0.863.

credit rating or age. Second, as a proxy for contemporaneous loan shares, our loan share measure is computed using average debt between periods t and t-1. This mitigates any contemporaneity concerns. Finally, loan shares and firm shocks are reassuringly uncorrelated in our sample[15].

To confirm the validity of the instrument, in Figure 2.3 we plot the relationship between the GIV and the raw endogenous covariate $\bar{\epsilon}_{i,t}$. There is a strong, positive relationship between the two variables with a Pearson correlation of 0.863.

Factor Analysis and Heterogeneous Loadings In Section 2.9.2 of the Online Appendix we study an important extension of the baseline approach that relaxes the assumption of homogenous loadings on bank factors. In other words, we allow $\delta_{i,j}$ to be *heterogeneous* across firms. In principle, sensitivity to fluctuations of bank x year factors could vary significantly across the firm distribution. We thus run a series of Principal Component Analyses (PCA) on each bank's portfolio separately and extract common components nonparametrically. We will find that all main results remain quantitatively unchanged.

[15]The raw correlation between loan shares and firm shocks in our sample is –0.02. The correlation is computed for each bank, and we report the average across banks.

2.3.5 Granular Credit Risk Spillovers: Credit Supply and Firms

In order to study the economic consequences of granular credit risk, we investigate the relationship between bank-level aggregated firm shocks and credit market outcomes. We follow a large literature in banking relying on the methodology in Khwaja and Mian (2008). Specifically, we focus on a sub-sample of firms borrowing from multiple banks and compare - for the same firm - quantity and rate outcomes from banks that experienced good or bad granular credit shocks. We test whether banks pass on shocks originating from their granular borrowers to the rest of their credit portfolio (non-granular borrowers). We define non-granular borrowers as firms whose loan share is below a certain threshold (such as the 50%th or the 25%th percentile) of the loan share distribution. In response to negative shocks due to granular clients, bank may have to scale back their relationship with non-granular borrowers, alter the pricing of loans, or both.

We run the following regressions on yearly changes:

$$\Delta y_{i,j,t} = \alpha_i + \alpha_{j,t,s(j),z(j)} + \beta \Delta \hat{u}_{i,t} + \nu_{i,j,t} \tag{2.9}$$

where $\Delta \hat{u}_{i,t}$ is the fitted value from the "first-stage" regression of the endogenous bank level shock $\Delta \bar{\epsilon}_{i,t}$ on the granular instrument $\Delta GIV_{i,t}$, $\alpha_{j,t,s(j),z(j)}$ is a firm x year x industry x county fixed effect and α_i is a bank fixed effect. $\Delta y_{i,j,t}$ is either loan volumes or interest flows. The regression is run either on firms with a loan share below the 50^{th} or 25^{th} percentile of the loan share distribution.

After investigating loan-level responses, we aggregate our data to the firm level and test whether there are any spillover effects from granular credit shocks to firm balance sheet aggregates such as investment or cash balances. We also look at the impact of granular credit risk on firm bankruptcies, contemporaneously or with a lag. We run the following firm-level regressions:

$$\Delta y_{j,t} = \alpha_{s,z,t} + \beta \Delta \hat{u}_{j,t} + \nu_{j,t} \tag{2.10}$$

with similar notations as above. $\Delta y_{j,t}$ are now firm-level outcomes such as (changes in) capital, sales, wage bill, cash as well as probability of bankruptcy (in levels). Essentially, in these spillover regressions the bank × year series of GIV-instrumented firm shocks is treated as a typical liquidity shock to the intermediaries' balance sheet, which is then passed on to the rest of the economy as a supply-side disturbance. The difference between our paper and the rest of the literature is that the origin of this bank-side liquidity risk is (uninsurable) idiosyncratic risk from large, granular borrowers[16].

[16]We test the insurability of granular credit risk in Section 2.4.3.

Table 2.2: **Loan Outcomes**

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Return on Loan (RoL)				
Firm Shock	0.134	0.146	0.334	0.335	0.361
	(0.012)	(0.013)	(0.015)	(0.017)	(0.019)
Bank x Industry x Year FE	No	No	No	Yes	No
Bank x Industry x Year x Loan-type x County FE	No	Yes	No	No	Yes
Additional controls in equation 2.1	No	No	Yes	Yes	Yes
Observations	479754	434662	333289	317186	292825
R^2	0.000	0.195	0.001	0.114	0.167
$\mathbb{E}$(RoL)	7.988%	7.975%	9.012%	9.029%	9.098%
SD(RoL)	7.993%	7.928%	8.921%	8.928%	8.923%

Notes: This table reports results from the regression of loan-level returns on loans on idiosyncratic firm shocks. The exact specification is described by equation (2.2). Columns (1)-(2) are based on firm shocks from specifications where (log) value-added is regressed only on a set of firm and county×industry×year×loan type fixed effects. Columns (3)-(5) are based on specifications that include additional firm controls: total assets, wages, leverage, liquidity, credit rating, age, and age squared. Firm shocks are normalized by their standard deviation. For example, 0.334 should be interpreted as an increase in the return on loans of 33.4 basis points in response to a 1 standard deviation idiosyncratic firm shock. Loan types include regular and credit-line loans. Counties are 19 administrative areas (*fylke*) in Norway. Industries are 99 2-digit sectors. Standard errors (in parentheses) are double clustered at the firm-year level and corrected for the estimated regressor bias with bootstrapping. The last two rows report the unconditional sample mean and standard deviation of the dependent variable.

2.4 Main Empirical Results

We investigate how idiosyncratic firm value-added shocks affect loan returns in section 2.4.1. In section 2.4.2, we aggregate idiosyncratic firm shocks to the bank level and see whether the effect is still significant despite portfolio-level risk pooling. In section 2.4.3 we ask whether granular credit risk goes unhedged at the bank level. In section 2.4.4, we investigate potential spillover effects from granular credit risk onto other firms and their real economic consequences.

2.4.1 Loan Outcomes

Table 2.2 presents the effect of idiosyncratic firm shocks on loan returns obtained from estimating Equation (2.2). Overall, idiosyncratic firm shocks have a large and significant (at the 1% level) effect on loan-level returns. In columns (1)-(2) we proxy firm shocks with the residual value-added variation after controlling for fixed effects only (as in di Giovanni et al. (2014)).[17] In columns

[17]This specification has no additional controls and extracts firm-level value-added variation that is orthogonal to industry × county × time × loan type and firm fixed effects.

(3)-(5) the shock is extracted from Equation 2.1 with a full set of firm controls: leverage, liquidity, size, age, and credit rating. Firm shock measures are standardized. Our preferred specification is column (5) and the result is the following: a 1-standard-deviation firm shock affects returns by 36.1 basis points. In words, when comparing a bank's loan return across firms within the same year, industry, county, and through the same loan facility, a 1 standard deviation reduction in firm performance reduces loan returns by roughly a third of a percentage point.

Figure 2.5 of the Online Appendix reports loan outcome estimates at different horizons: we regress loan returns on leads and lags (in years) around the firm shock ("event" at date 0) and plot the dynamic of the interval estimates over time. First, we observe that there is no effect for years prior to the shock, which points at the absence of any pre-trends. Second, the impact of idiosyncratic firm shocks on loan outcomes is felt for a long time: at least for 3 years on average. We interpret this result through the lenses of relationship-based lending. Termination of a credit relationship is costly, for either side, because of the presence of asymmetric information in credit markets. Even if a bad idiosyncratic outcome reveals new information about the borrower's "type", ex-post monitoring of the repeated borrower may still be a more cost-effective alternative than forming a new relationship (Williamson, 1987). Lenders may understand and internalize the adverse selection problem in the market for switching borrowers (Sharpe, 1990). Finally, the cost of asymmetric information may be bigger for smaller firms, which are also potentially more likely to experience a negative idiosyncratic shock and have more to gain from sticking to the original lender (Chodorow-Reich, 2014). In equilibrium, the lender agrees that the borrower postpones a fraction of the loan repayment to the future period.[18]

2.4.2 Granular Credit Risk: Bank Outcomes

The finding that firm-level idiosyncratic shocks impact loan returns merely reflects the fact that individual loans are inherently risky investments. There is little margin of adjustment for the bank to insure against bad loan-level outcomes. The natural next question is whether these idiosyncratic shocks average out at the level of bank *portfolios*. In other words, can/do banks take advantage of risk pooling and diversify idiosyncratic firm risk away? To answer this question we proceed by estimating the relationship in (2.7). Results are reported in Table 2.3, where we have normalized the bank shock by its standard deviation.

We report two sets of specifications: with and without the granular instrumental variable (GIV). In the first two columns (OLS estimates) we find that even at the level of banks' portfolios, idiosyncratic credit risk is associated with large and significant effects on bank returns. To address

[18]Our extensive margin analysis in Section 2.5.1 will reveal that it is indeed the intensive margin, i.e. temporary non-performance and payment delay, which drives our loan-level results, and not necessarily firm exit.

Table 2.3: **Bank Outcomes**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable: Bank Return on Loans (RoA)							
	OLS		Instrumented with GIV					
	Pooled	Pooled	Pooled	Positive	Negative	Pooled	Positive	Negative
Granular Credit Shock	0.129	0.136	0.116	0.016	0.194	0.117	0.056	0.176
	(0.029)	(0.027)	(0.031)	(0.094)	(0.074)	(0.030)	(0.087)	(0.072)
First stage F-stat			1429.683	138.772	396.907	1137.722	150.136	263.982
J-statistic			0	0	0	0	0	0
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bank Controls	No	Yes	No	No	No	Yes	Yes	Yes
Observations	1211	1211	1211	508	694	1211	508	694
R^2	0.752	0.770	0.599	0.646	0.569	0.627	0.683	0.592
$\mathbb{E}$(RoA)	6.350%	6.350%	6.350%	6.460%	6.289%	6.350%	6.460%	6.289%
Sd(RoA)	1.354	1.354	1.354	1.403	1.295	1.354	1.403	1.295

Notes: This table reports results from regressing bank-level return on loans on bank-level aggregated firm shocks $\bar{\epsilon}_{i,t}$. Columns (1)-(2) are standard OLS on equation (2.3), while columns (3)-(8) instrument the aggregated shock by the granular IV as in equation (2.7). The GIV is constructed following equation (2.6). Positive (negative) shock specifications include only observations in which the bank shock $\bar{\epsilon}_{i,t}$ is above (below) zero. Bank controls include lagged total assets, leverage, liquidity, number of loans, deposit to assets ratio and financial assets to assets ratio. The last two rows report the unconditional sample mean and standard deviation of the dependent variable. The F-stat is the Kleibergen-Paap rk Wald F statistic for the test of weak identification. J-stat is the Hansen's J-statistic for the instrument overidentification test. Standard errors (in parentheses) are clustered at the bank level and corrected for the estimated regressor bias with bootstrapping.

potential endogeneity concerns, columns (3)-(8) report results from the IV regression[19]. Our results show that a one standard deviation GIV-instrumented firm shock, on average, affects bank loan portfolio returns by 11-12bps. We have specifications with and without additional bank controls which include lagged values of book leverage, liquidity, total assets, number of loans, deposit-to-asset ratio, and financial assets to total asset ratio[20]. Results are qualitatively and quantitatively robust to the exclusion of these controls[21].

[19]Formal statistical diagnostic tests show validity for the GIV as a good instrument. The first-stage F-statistic is above the Stock and Yogo (2005) criterion for 5% maximal relative bias. The Hansen J-statistic cannot reject the null hypothesis of the instrument being exogenous.

[20]Theoretically, if the exclusion restriction holds, the GIV approach would not require any further bank-time controls. The reason is that GIV, by construction, would be purged from any bank-time factors. For robustness, we still include observable bank controls. Results do not change in any substantial matter, which adds validity to the method. In addition, in Section 2.9.2 we also control for latent bank-time factors, extracted using PCA. Results do not change.

[21]Bank-level return on corporate loans is the main dependent variable in this section. We have also experimented

In Figure 2.7 of the Online Appendix we report bank loan outcomes by horizon. We find that the impact of GIV-instrumented firm shocks on bank RoA lasts for up to 1 year, i.e. a shock at t has a significant effect on returns even at t+1. In addition, the effects of lags are not significant implying the absence of any pre-trends.

A second key set of results is related to the asymmetric effects of granular firm risk. In columns (4)-(5) and (7)-(8) of Table 2.3 we explore positive- and negative-only firm shocks with and without bank controls. Specifically, we condition on the endogenous covariate $\bar{\epsilon}_{i,t}$ in equation (2.3) being positive or negative only, and instrument it by the GIV. Only negative shocks have a significant impact on bank returns. The impact of positive shocks is not statistically significantly different from zero. A one standard deviation negative granular firm shock lowers bank returns by up to 19.4bps, which is much larger than the average effect and amounts to roughly 15% percent of the standard deviation of banks' portfolio returns. Due to the payoff structure of the debt contract this very concave relationship is not surprising. Because of debt contracts, banks find it difficult to extract higher dividends from firms that are performing well, while at the same time remaining exposed to potential downside risk from firms that perform poorly. In case of a negative shock, the firm's loan may become nonperforming, the firm may default on the obligation, or exit the industry altogether[22].

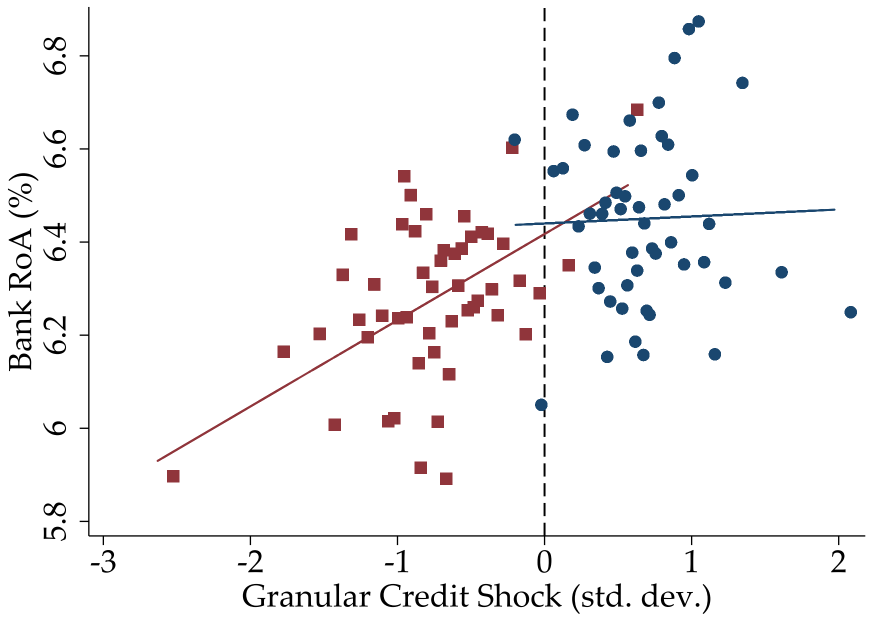
Figure 2.4 provides a visual representation of this concave relationship. The figure depicts the (binned) scatter plot of the impact of GIV-instrumented firm shocks on banks' returns on loans (RoA). Blue circles (red squares) represent positive and negative shocks, respectively. We construct the binned scatter plots by first regressing both bank RoA and the GIV-instrumented firm shocks on bank and time FE, computing the residuals, and adding back the mean of each variable. We then construct 50 equally-sized bins of the residual shock variable. Figure 2.4 plots the mean residual bank RoA within each bin versus the bin's mean residual shock. Finally, we overlay the linear fits for the respective specifications. The asymmetry of the result is rather striking: the line of best fit for positive shocks is flat, while the slope for negative shocks is downward-sloping and highly significant. The bins are all equally-sized, so each dot represents 10+ underlying bank $\times$ time observations. Our results are thus not driven by any individual outliers. We interpret the concave relationship as further validation that our measure of firm shocks is indeed economically informative.

---

with loan writedowns and portfolio-level Sharpe ratios. Table 2.19 of the Online Appendix reports the results. We find some evidence that granular credit risk, when instrumented by the GIV, is weakly positively (negatively) associated with the Sharpe Ratio (writedowns)

[22]We explore the extensive margin in detail in Section 2.5.1.

Figure 2.4: **Granular Credit Risk and Bank Outcomes**



Notes: This figure visualises the relationship between residualized bank-level return on loans and residualized instrumented bank-level aggregated firm shocks. The red squares (blue circles) are binned scatterplots conditional on negative (positive) values of the weighted firm shock $\bar{\epsilon}_{i,t}$. The shock variable is normalized by its standard deviation. We construct the conditional binned scatterplot in three steps, and each step is performed separately on positive and negative values of $\bar{\epsilon}_{i,t}$. First, we residualize bank-level returns on loans and instrumented firm shocks. Instrumented shocks represent fitted values from regressing $\bar{\epsilon}_{i,t}$ on the GIV. The residualized return and shock values are obtained from regressing each variable on bank and time fixed effects, computing the residual, and adding back the mean of each variable. Second, we construct 50 equally-sized bins based on the residualized shock. Third, we plot the mean residual bank return within each bin versus the bin's mean residual shock. The red (blue) line represents the linear fit from regressing bank-level loan return on instrumented shocks, conditional on $\bar{\epsilon}_{i,t} < 0 (> 0)$.

2.4.3 Hedging

We have so far established that idiosyncratic shocks to individual corporate clients affect bank portfolio returns. However, it is possible that financial intermediaries hedge granular credit risk with derivatives and other instruments. As a first pass attempt in answering this question, we collect bank-level data on income from fees, derivatives, equity and bond holdings, and dividends. We then correlate changes in returns from these sources with our GIV-instrumented shocks. The conjecture is that in the same state of the world in which banks are hit with bad idiosyncratic shocks to their loan books, returns are compensated through alternative departments within the same bank. For example, banks could command higher fees for late interest payments, hedge negative states with credit derivatives, short stocks of firms they are also lending to, etc. Table 2.4 reports the results.

As can be seen from the table, the data cannot consistently reject the null hypothesis of little to no insurance against granular credit risk. None of the measures of non-interest income are

Table 2.4: **Hedging Granular Credit Risk**

	(1)	(2)	(3)	(4)	(5)
Dependent Variable: Δ Income from	Fees	Derivatives	Equity	Bonds	Dividends
			Pooled		
Granular Credit Shock	0.219	-0.658	-1.323	0.163	0.173
	(0.131)	(1.214)	(1.477)	(0.140)	(0.631)
Bank FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	1211	344	1058	1197	1174
R^2	0.010	0.049	0.011	0.013	0.046
			Negative Shocks Only		
Granular Credit Shock	0.330	-0.133	-3.420	0.461	-0.209
	(0.236)	(2.944)	(5.466)	(0.470)	(0.170)
Bank FE	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes
Observations	697	197	606	690	680
R^2	0.021	0.037	0.023	0.021	0.164

Notes: This table reports results from regressing bank-level year-on-year growth rate in non-interest income components on bank-level aggregated firm shocks, instrumented by the granular IV. The top panel presents results for all shocks, positive or negative. The bottom panel presents results for negative shocks only ($\bar{\epsilon}_{i,t} < 0$). The granular IV is constructed based on equation (2.6). Standard errors (in parentheses) are clustered at the bank level. Data on all bank non-interest income is from the financial supervisory database ORBOF.

significantly associated with our shock measure. If anything, some forms of income are in fact *positively* correlated with idiosyncratic credit shocks, which questions their usefulness as a hedging instrument.

A drawback of this analysis is that the various hedging instruments analyzed in Table 2.4 are only observable at the bank level. A more detailed analysis would construct matched derivatives holdings at the level of individual credit relationships. This would increase the odds for banks to hedge *firm-specific* risk, something that we can not plausibly establish by looking at portfolio-level data. This would be possible only for a very small subset of large firms that are (a) listed and (b) have a liquid market for credit derivatives such as credit default swaps (CDS). The mass of such firms is small and the CDS market is not very liquid in Norway. Regardless, insurability of granular credit risk is an important question, to which we can give only a partial answer given the data constraints.[23]

[23]Banks could also dilute single-name concentration risk by engaging in syndicated lending. In the case of Norway, however, syndicated loans constitute a very small fraction of external financing for firms.

2.4.4 Granular Credit Risk Spillovers: Credit Supply and Firm Outcomes

Previous sections have documented that granular credit risk has quantitatively important effects on bank portfolio outcomes, and that this risk is unhedged. In this section, we ask whether banks hedge these shocks "ex-post", i.e. by passing it on to the rest of their corporate portfolio. We are interested in seeing whether banks react by reducing loan supply or raising interest rates, in particular for non-granular firms.

Table 2.5 reports our results on the supply of credit. Our approach follows closely Khwaja and Mian (2008). In all specifications we impose a stringent configuration of interacted firm × year × industry × county fixed effects. Our specifications regress year-on-year changes in the granular credit shock on year-on-year changes in loan-level credit supply. In columns (1)-(2), we look at the impact of bank-level granular credit shocks - either instrumented or not - on all firms and find no significant relationship. In columns (3)-(4) we restrict the sample to non-granular firms only. Non-granular firms are defined as those whose bank loan shares are below the 50^{th} (column (3)) or 25^{th} (column (4)) percentiles of the loan share distribution. For example, the median loan share is 0.000211 percent across all loans. We do find a statistically significant relationship in this case, particularly when the threshold for non-granular firms is the 25th percentile. In columns (5)-(6) we add a bank fixed effect to the baseline specification and results do not change substantially. Overall, a one-standard deviation decline in the granular credit shock reduces loan supply growth to non-granular borrowers by up to 71 basis points. This effect is strongly significant. These results are suggestive of a "pecking order" of credit relationships where banks keep credit relationships with their main clients unchanged but adjust lending conditions with their non-granular borrowers in order to compensate for portfolio losses.

In Table 2.6 we repeat the same exercise but with interest flow as the left-hand-side variable. We find a strong negative relationship between year-on-year changes in granular credit risk and yearly growth in loan-level interest flows. We interpret these changes in flows as an effect on the loan interest rate. A one-standard deviation decline in the granular credit shock increases interest rate growth on loans to non-granular clients by up to 63.4 basis points. Taken together with the positive association with credit quantities, we have identified granular credit risk as a textbook supply-side disturbance: a negative granular credit shock results in a leftward shift in the supply schedule, leading to a reduction in quantities and elevation in credit market prices. In addition, the pass-through mechanism can also be interpreted as operating through a kind of bank credit supply network: two firms that may otherwise not be connected can impact each other's performance through their association with a common lender[24].

[24]We control for firm county and industry fixed effects as well as the bank fixed effects. We cannot fully rule out, however, that these firms are not also associated through a production network as we do not have the data on firm linkages to test that hypothesis. We return to this point in Section 2.6.

Table 2.5: **Spillovers from Granular Credit Shocks: Credit Supply**

	(1)	(2)	(3)	(4)	(5)	(6)
Δ Bank Shock	0.023	0.022	0.165	0.625	0.168	0.717
	(0.043)	(0.043)	(0.129)	(0.288)	(0.136)	(0.311)
Year x Industry x County x Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Bank FE	No	No	No	No	Yes	Yes
Non-Granular Firms (50%)	No	No	Yes	No	Yes	No
Non-Granular Firms (25%)	No	No	No	Yes	No	Yes
Instrumented by GIV	No	Yes	Yes	Yes	Yes	Yes
Observations	15279	15279	3449	348	3413	322
R^2	0.484					

Notes: This table reports results from regressing year-on-year changes in (log) bank debt at the bank-firm level on the year-on-year change in bank-level aggregated firm shocks which are either instrumented by the granular IV as in columns (2)-(6) or not, as in column (1). Specifications are based on equation (2.9). The GIV is constructed based on equation (2.6). Columns (1)-(2) include all firms. Columns (3)-(6) include only non-granular firms. Non-granular firms are defined as firms whose bank loan shares are less than the 50th (columns (3)-(4)) or the 25th (columns (5)-(6)) percentiles of the loan share distribution, which is pooled over all banks and years. The full distribution of loan shares is plotted on Figure 2.1. Standard errors (in parentheses) are double clustered at the bank and firm level.

Next, we ask whether the impact on non-granular firms ultimately leads to significant economic consequences. We aggregate our data to the firm level and consider several firm outcomes as dependent variables. Those include growth in sales, the wage bill, capital and the cash position. In all of the specifications, we include year × industry × credit rating, as well as firm and bank fixed effects. In addition, we focus on the samples of non-granular firms where non-granular firms are defined as those whose bank loan shares are below the 50^{th} or 25^{th} percentile of the global distribution of loan shares. In other words, we trace out the economic consequences of a credit supply shock on the same non-granular firms that we show were impacted in Tables 2.5 and 2.6.

Results are reported in Table 2.7. The most interesting results are in columns (2)-(3), where we document that a change in the granular credit shock at the bank level is positively associated with capital growth at the firm level for non-granular borrowers. This suggests that a one-standard deviation negative credit supply shock causes a decline in firms' fixed capital investment growth by roughly 13-24 basis points. The impact on non-granular firms defined by the median loan share cut-off is strongly statistically significant. This result has immediate implications for the real macroeconomy. We also considered outcomes such as sales, the wage bill and the firms cash position. For these variables our coefficient estimates are imprecise, and in all of these cases we are unable to reject the null hypothesis.

Even if non-granular firms have relatively low loan shares, they constitute an important fraction of the economy. Specifically, non-granular firms that are defined by the median loan share cutoff

Table 2.6: **Spillovers from Granular Credit Shocks: Interest Rates**

	(1)	(2)	(3)	(4)	(5)	(6)
Δ Bank Shock	-0.004	-0.017	-0.361	-0.341	-0.421	-0.634
	(0.064)	(0.066)	(0.189)	(0.417)	(0.190)	(0.448)
Year x Industry x County x Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Bank FE	No	No	No	No	Yes	Yes
Non-Granular Firms (50%)	No	No	Yes	No	Yes	No
Non-Granular Firms (25%)	No	No	No	Yes	No	Yes
Instrumented by GIV	No	Yes	Yes	Yes	Yes	Yes
Observations	15279	15279	3449	348	3413	322
R^2	0.533					

Notes: This table reports results from regressing year-on-year changes in (log) interest flows at the bank-firm level on the year-on-year change in bank-level aggregated firm shocks, which are either instrumented by the granular IV (columns (2)-(6)) or not (column (1)). Specifications are based on equation (2.9). The granular instrument is constructed based on equation (2.6). Columns (1)-(2) includes all firms. Columns (3)-(6) include only non-granular firms. Non-granular firms are defined as firms whose bank loan shares are less than the 50th (columns (3)-(4)) or the 25th (columns (5)-(6)) percentiles of the loan share distribution, which is pooled over all banks and years. The full distribution of loan shares is plotted on Figure 2.1. Standard errors (in parentheses) are double clustered at the bank and firm level.

account for at least 15% of total aggregate capital in the sample in any given year. Granular credit shocks therefore affect a sizable fraction of aggregate capital via credit spillovers. It is also important to emphasize that our estimates constitute a *lower bound* - we can not rule out further second- and third-order spillover effects to granular clients as well. Granular credit risk can therefore potentially affect an even larger fraction of aggregate investment through the broader credit supply network as well as through general equilibrium effects.

Finally, we investigate whether the reduction in credit not only affects firm balance sheet outcomes but ultimately firm default. We ask whether a change in the granular credit shock at the bank level is associated with a higher frequency of bankruptcies at the firm level. Table 2.8 reports the results from our baseline probit regressions. In column (1), we include all firms in the estimation sample. In column (2), we restrict the sample to non-granular firms only, with the threshold being the 50^{th} percentile of the loan share distribution. For both specifications we find a strong negative association. For columns (4)-(6) and (7) we trace out the impact of granular credit shocks in t-1 and t-3, respectively, on the probability of bankruptcy at t. In columns (3) and (6) we run the same regressions but restricting the sample to firms with a loan share below the 25^{th} percentile of the distribution. Finally, in column (8) we regress the probability of a firm filing for bankruptcy at any point over its existence in our dataset on its average granular credit shock. That is, we ask if firms that ever default also experience, on average, worse granular credit supply shocks from their lenders. Overall, across all 8 specifications, we find a very strong negative association

Table 2.7: **Firm Outcomes from Granular Credit Shocks**

	(1) Capital	(2) Capital	(3) Capital	(4) Sales	(5) Wages	(6) Cash
Δ Bank shock	0.040	0.241	0.129	0.001	0.007	0.142
	(0.030)	(0.095)	(0.251)	(0.031)	(0.040)	(0.146)
$\mathbb{E}$(dependent variable)	-0.081	-0.095	-0.105	0.026	0.034	0.067
Sd(dependent variable)	0.603	0.640	0.683	0.290	0.344	1.037
Year $\times$ Industry $\times$ County FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Non-Granular Firms (50%)	No	Yes	No	Yes	Yes	Yes
Non-Granular Firms (25%)	No	No	Yes	No	No	No
Instrumented by GIV	Yes	Yes	Yes	Yes	Yes	Yes
Observations	90800	39861	15444	44547	45452	43994

Notes: This table reports results from firm-level regressions where the outcome variable is either firm-level year-on-year change in log(capital), log(sales), log(wage bill), or log(cash). The control variable is the year-on-year change in bank-level aggregated firm shocks which are instrumented by the granular IV. Specifications are based on equation (2.10). The granular instrument is constructed based on equation (2.6). Non-granular firms are defined as firms whose bank loan shares are less than the 50th or 25th percentile of the loan share distribution. For firms with multiple banking relationships, we define a firm as non-granular if the largest loan share of that firm across all credit relationships is less than the 50th or the 25th percentile of the loan share distribution. For these firms, the bank shock is computed as the average across all lending relationships. The full distribution of loan shares is plotted on Figure 2.1. All specifications include interacted year x industry x county fixed effects and firm fixed effects. Standard errors (in parentheses) are clustered at the firm level.

between granular credit risk and firm bankruptcy probability. The impact is quantitatively very large - a one-standard deviation negative granular credit supply shock increases the likelihood of bankruptcy by roughly 32-60 basis points for all firms, and 68-106 basis points for non-granular borrowers. Additionally, we do find that firms that went bankrupt at some point in the sample are also those that experienced abnormally bad granular credit supply shocks. Our results have direct implications for the aggregate economy and consumer welfare, considering that firm bankruptcy proceedings are very costly in practice.

We conclude this section by reiterating our main findings. First, idiosyncratic firm shocks have a large, long-lasting and significant effect on loan-level returns. Second, these shocks survive portfolio aggregation and impact bank-level outcomes. Importantly, these shocks originate from granular, i.e. large, borrowers which is precisely the reason why they do not wash out. Third, banks do not hedge granular credit risk with income from non-loan businesses such as derivatives or equity investments. Fourth, there are considerable loan-level spillovers of granular credit shocks on non-granular borrowers: affected banks reduce loan supply and increase interest rates on their less important, non-granular clients. Fifth, those affected clients in turn reduce their investment

Table 2.8: **Firm Bankruptcy from Granular Credit Shocks**

Probit Model	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Probability of bankruptcy$_t$				Ever bankrupt
$\Delta\text{BankShock}_t$	-0.609	-0.680	-1.056					
	(0.110)	(0.196)	(0.307)					
$\Delta\text{BankShock}_{t-1}$				-0.322	-0.965	-0.946		
				(0.123)	(0.203)	(0.334)		
$\Delta\text{BankShock}_{t-3}$							-0.703	
							(0.239)	
$\Delta\text{BankShock}_t$								-1.273
								(0.281)
Non-Granular Firms (50%)	No	Yes	No	No	Yes	No	Yes	Yes
Non-Granular Firms (25%)	No	No	Yes	No	No	Yes	No	No
Instrumented by GIV	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	61819	35965	20161	50897	29451	16302	16648	35965

Notes: This table reports results from firm probit regressions of likelihood of firm bankruptcy on the bank-level granular credit shock. In columns (1)-(7), the outcome variable is probability of contemporaneous firm bankruptcy. In column (8), the outcome variable is the probability that a firm ever goes bankrupt. In columns (1)-(3), the control variable is the contemporaneous mean (change in the) bank-level credit shock which is instrumented by the granular IV. In columns (4)-(7), the RHS variable is the GIV-instrumented bank-level credit shock lagged by either one or three years. Columns (1) and (4) are for all firms in sample. Remaining columns restrict the sample to non-granular firms only. Non-granular firms are defined as firms whose bank loan shares are less than the 50th or the 25th percentiles of the loan share distribution, which is pooled over all banks and years. For firms with multiple banking relationships, we define a firm as non-granular if the largest loan share of that firm across all credit relationships is less than the 50th or the 25th percentile of the loan share distribution. For these firms, the bank shock is computed as the average across all lending relationships. The full distribution of loan shares is plotted on Figure 2.1. Standard errors in parentheses. Firm bankruptcy information is from the credit rating agency Bisnode.

in physical capital and are much more likely to file for bankruptcy. Overall, our results show that idiosyncratic shocks to granular borrowers have important implications for the broader financial and real economy. In the language of financial regulators, single-name credit concentration risk is quantitatively important.

2.5 Heterogeneity and Robustness

In this section, we first provide further evidence on the underlying mechanisms behind our findings. In 2.5.1 we focus on whether certain firms are more likely to transmit idiosyncratic firm performance shock to their banks and in 2.5.2 whether certain banks are more exposed to granular credit risk. In 2.5.3 we discuss several important robustness checks.

2.5.1 Firm Heterogeneity

Loan Outcomes - Balance Sheet Heterogeneity We start by exploring heterogeneous effects of idiosyncratic firm shocks originating from firms with different characteristics. Specifically, we augment specification (2.2) by interacting our extracted shocks with lagged firm characteristics. We are interested in how the transmission mechanism differs for firms with high leverage, low asset size, low equity, short average debt duration, high bank credit reliance, low credit rating, and young age. Each characteristic is thus a dummy which equals 1 for firms in that particular category of interest and 0 otherwise.

Table 2.15 in the Online Appendix presents the results. There is overall rich firm heterogeneity behind our loan-level outcomes. Relative to the baseline, the pass-through of idiosyncratic firm shocks is stronger for firms with high leverage, low assets, low equity, short debt duration, high reliance on bank debt, lower-than-"A" credit ratings, and firms younger than 3 years. All of these firms, relative to the average firm, are more likely to be more "risky" from the bank's perspective. Interestingly, we find that interactions with firm size and debt duration are statistically different from other characteristics. For macro-prudential purposes, these results offer a new dimension for regulation of concentration risk: banks which are heavily exposed to, for example, small, risky, young firms are at much greater risk of suffering from detrimental idiosyncratic credit shocks than intermediaries that lend to liquid and non-levered corporates.

Loan Outcomes - Extensive Margin Are our loan-level results driven by the intensive or the extensive margin? We are interested in seeing whether the transmission of idiosyncratic firm shocks is different among firms that enter/exit the industry or go bankrupt. Our strategy is to construct a dummy variable for each of the three groups of firms. For entrants, the dummy takes the value of unity in the year following the entry, while for leavers and bankrupt firms the variable equals unity in the year prior to the event. We also consider an "ever-bankrupt" dummy which takes the value of unity for firms that filed for bankruptcy at any point during the 2003-2015 period. The latter variable captures potentially some unobserved intangible characteristics such as poor management skills, which are common for unsuccessful firms but cannot credibly be inferred from balance sheet information.

Table 2.16 in the Online Appendix reports the results. We see that the shock transmission mechanism is stronger (weaker) among firms which have just entered (about to exit) the industry. We do not find that the channel is stronger among firms which go bankrupt. Overall, the extensive margin is active but does not dominate our results. In other words, even conditional on firms being non-entrants, non-leavers, and not in bankruptcy, negative idiosyncratic shocks can cause lower bank returns. That implies that our results are driven mostly by the intensive margin.

Loan Outcomes - Ownership and Industry Heterogeneity Next, we investigate whether our results are driven by firms with a particular ownership structure or industry affiliation. For example, is the shock transmission stronger among special financial vehicles or construction firms? In Table 2.17 of the Online Appendix we report firm ownership heterogeneity results, along with our baseline estimates. We see clearly that our results reflect conventional privately owned firms and not state, community, or special financial vehicles. Privately owned firms dominate our sample by a wide margin.

Table 2.18 explores heterogeneous effects by firm sector. Our baseline estimates are almost identical to results from manufacturing firms. Overall, there doesn't appear to be any abnormality across different industries; the real estate sector is the only one where pass-through appears to be significantly smaller.

Loan Outcomes - Geographical Heterogeneity Are our loan-level results driven by idiosyncratic shocks to firms located in particular geographical regions of Norway? Figure 2.6 in the Online Appendix plots a coloured map of Norway, where each of the 19 counties is colored with a different shade of blue. Darker regions represent a higher local pass-through coefficient of idiosyncratic firm shocks onto loan-level returns. Recall that our baseline average pass-through estimate at the loan level is 0.361. Based on the map we document two main results. First, there is interesting cross-regional heterogeneity in the estimates that is potentially worth exploring in future research. Second, this heterogeneity is not too drastic: county-wide averages are in the [0.19,0.44] range[25]. Finally, we see that our result is not driven solely by Oslo and neighboring counties but is in fact present throughout the country. We therefore conclude that our results are likely not driven by some unusual regional clustering of correlated idiosyncratic shocks.

2.5.2 Bank Heterogeneity

In this section, we explore whether various banks are affected differentially by granular credit shocks aggregated at the portfolio level. We explore several dimensions of bank heterogeneity: portfolio risk weights, (log of) risk-weighted assets (RWA), regulatory capital ratio, loan portfolio Herfindahl (HHI), (log of) number of loans, liquidity ratio, and profitability ratio.[26] We compute portfolio risk weights by dividing RWA by book assets. The regulatory capital ratio is defined as regulatory capital over RWA. Liquidity is defined as the ratio of cash holdings to book assets.

[25]The exception is the northernmost county, Finnmark, where we find a point estimate of -0.10. However, this county is also by far the least populated area of Norway.

[26]We use RWA as a proxy for bank size, broadly defined. We have also experimented with book assets, book equity, and regulatory capital as alternative size measures. Results do not change. In addition, we also condition on whether banks are domestically or foreign owned. Baseline results are quantitatively very close to the sub-sample of privately-owned banks; estimates based on foreign banks are consistently imprecise.

Profitability is defined as the ratio of profits before taxes to book assets. All characteristics are lagged. For each characteristic we define a dummy variable based on the median of the respective lagged distribution. Table 2.20 of the Online Appendix presents the results. Each column reports coefficients for interactions of GIV-instrumented bank-level firm shocks and dummies for respective lagged bank characteristics. All specifications include the time and bank fixed effects as well as the usual set of bank controls.

From the table we observe several notable results. First, the number of loans does not materially affect the transmission of granular credit shocks, since the pass-through is significant for banks with a high number of loans (column (5)). This suggests that granular credit risk is not merely a "small N" problem. Second, the pass-through is stronger for banks with low RWA (column (2)) and high capital ratios (column (3)). The two effects are interconnected, since in the cross-section larger banks are more levered and thus have lower capital ratios.[27] Third, the pass-through is twice as large for banks with high loan portfolio concentration (column (4)). This is reassuring, since given the same volatility of idiosyncratic firm shocks, higher concentration should make banks more exposed to shocks stemming from the right tail of the loan share distribution. Last but not least, in column (1) we see that banks with higher risk weights are more exposed to granular credit shocks.[28] This is potentially an important finding because credit concentration risk and the risk-taking channel may form complementarities that could impact anything from the financial boom-and-bust cycle to the transmission of monetary policy (Bruno and Shin, 2015).[29] In order to inspect and better understand this mechanism, we now perform an additional exercise below.

We examine the impact of granular credit risk on bank returns, interacting the granular credit risk shock with both portfolio risk weights and other bank characteristics. Table 2.21 in the Online Appendix reports the results. Overall, we observe that (with the exception of the left column in specification (6)) estimates for banks with high risk weights are strikingly always higher than for banks with low risk weights. This strongly suggests that the credit concentration risk and risk-taking channels are positively associated. The most notable are results in columns (1), (3), and (6). These suggest that the pass-through of granular credit shocks, conditional on the sample of banks with high risk weights, is strongest if banks are small, have concentrated loan portfolios, and record high profits. The result on profits (column (6)) is particularly intriguing since it is consistent

[27]The observation that smaller banks are more exposed to granular credit shocks is in line with the existing theories that emphasize the role of bank size heterogeneity in the transmission of aggregate and idiosyncratic disturbances (Stavrakeva, 2019; Davila and Walther, 2020; Jamilov, 2021). In particular, smaller banks tend to have a greater balance sheet sensitivity with respect to exogenous shocks.

[28]Risk weights are not correlated with any of the proxies of bank size: RWA, capital, book assets, or book equity. They are also uncorrelated with the bank-level share of corporate credit to total assets.

[29]It is possible that banks with high risk weights are exposed to firms that are inherently riskier, in the spirit of (Chang et al., 2021). Our firm balance sheet heterogeneity analysis in Table 2.15 concluded that returns on loans that are extended to riskier firms are more likely to be affected by idiosyncratic firm shocks.

with the risk-taking channel: in good states of the world, i.e. when individual firm performance is high, banks with low risk aversion build riskier, concentrated portfolios and record higher profits. However, as our paper argues, this comes at the (potentially unhedged) cost of greater exposure to granular credit risk and eventual portfolio losses during the bad state, i.e. when firm performance is low. Overall, our results add an interesting new angle of portfolio concentration to the literature on endogenous financial cycles driven by risk taking of financial intermediaries (Coimbra and Rey, 2019).

2.5.3 Robustness Checks

In this section, we provide several additional robustness checks. We show that our results are robust to the Great Financial Crisis (GFC). We check that idiosyncratic firm shocks have a pairwise correlation of approximately zero. Finally, we conduct two placebo tests at various levels of aggregation to lend further support to our baseline results.

Robustness to GFC In order to investigate whether the relationship between granular credit risk and loan or bank outcomes is robust to the Great Financial Crisis, we redo our estimation focusing on years before or after the GFC. Table 2.22 in the Online Appendix reports the results. We highlight three main observations. First, our results do not vanish for either of the two sub-periods. Second, this is true for both loan-level and bank-level estimations. Third, estimates are slightly noisier for the pre-GFC period, although still statistically significant.

Pairwise Correlations Test An important question that must be addressed is potential piece-wise correlation of our idiosyncratic firm shocks. Systematic residual correlation across firms may indicate that our shocks are still driven by common factors, which would invalidate our conjecture that fluctuations are truly idiosyncratic. For example, we could be capturing some unobserved network effects such as the ones induced by firm trade credit relationships. To test this, we compute pairwise correlation coefficients across any two pairs of firms in our sample. Figure 2.9 of the Online Appendix presents the result. In total we have 1,861,485 pairs across a balanced panel sub-sample of our data. The average pairwise correlation is 0.019 and the standard deviation is 0.342. This implies that the average correlation coefficient can not be statistically distinguished from zero. This provides reassuring evidence in support of our idiosyncratic firm shocks being truly idiosyncratic and not being driven by unobserved factors that induce cross-sectional correlation, such as production networks.

Placebo Regressions To ensure that we do not falsely reject the null hypothesis due to potentially serially correlated error-terms, we follow Chetty et al. (2009) and implement a nonparametric permutation test for whether the true effect of idiosyncratic firm shocks on loan returns is zero. In order to do so, we randomly reassign the estimated firm-level shocks and redo the analysis at the loan and bank levels. Placebo Monte-Carlo permutations results are reported in Table 2.23 of the Online Appendix. We find that we can reject the null hypothesis of no association (at the 1% level) under this non-parametric distribution. In words, it's highly unlikely that our results are due to random chance. Furthermore, at the level of the bank, we confirm that our finding of strong asymmetric effects is not coincidental since the permuted positive-only shock estimate has a p-value of 0.82, while the negative-only shock estimate has a p-value of 0.000.

Further, to illustrate how our idiosyncratic shocks pick up economically meaningful information, we run a series of placebo regressions where firm shocks are randomly drawn from a uniform distribution instead of being extracted from the economic specification 2.1. The results from using these drawn shocks at the loan- and bank-level are reported in Table 2.24 of the Online Appendix. Across all specifications and levels of aggregation we find no association between these random shocks and loan or bank outcomes.

Fitting a Fixed Effects Model with AR(1) Disturbance We run our firm shocks through an autoregressive linear model of order 1 in order to establish whether shocks are persistent or not. We also want to facilitate future structural analysis of models with a financial sector that is subject to "idiosyncratic granular borrower risk". Specifically, we fit the full cross-section of firm shocks into a linear fixed effects model with an AR(1) disturbance term. Results are reported in Table 2.25 of the Online Appendix. Parameters of the process - autoregressive coefficient and standard deviation of the error term - are reported for all levels of aggregation. Overall, we find that the idiosyncratic firm shock is volatile (standard deviation of roughly 0.2) and not persistent at all (auregressive coefficient of roughly 0.12-0.32). A volatile iid process is likely to approximate granular credit risk rather well.

2.6 Discussion

In this section, we discuss several issues that are relevant for our empirical analysis. First, we discuss the literature on firm and bank credit networks and its implications for our results and methodology. Second, we discuss potential *origins* of large exposures.

2.6.1 Network Effects

A vibrant new literature emphasizes the role of bank and firm credit networks in the amplification and propagation of non-systematic shocks. One stream of the literature shows that bank credit supply shocks propagate along firm production networks, causing sizable real economic effects (Huremovic et al., 2020; Dewachter et al., 2020). We relate to this literature in two ways. First, in our analysis of economic spillovers, we essentially identify a *credit supply network*: there exists a pass-through mechanism for idiosyncratic borrower risk between two potentially unrelated firms that are "connected" via the balance sheet of their common lender. Second, a negative granular credit shock, passed onto non-granular borrowers via higher rates or lower quantities of credit, is a leftward shift of the credit supply curve and maps into the bank credit supply shocks in Huremovic et al. (2020) and Dewachter et al. (2020). We establish the first-degree direct effect of these shocks on firm performance and bankruptcy rates. However, because we do not observe the Norwegian input-output or firm trade-credit network, we can not speak of higher-degree effects from these networks. Thus, our analysis of economic spillovers of granular credit shocks in Section 2.4.4 establishes a *lower bound* on the total pass-through to the real economy.

Another stream of the literature such as Elliott et al. (2020) has recently shown that when banks are individually exposed to the same set of firm industries, idiosyncratic industry-specific risk gets amplified as opposed to getting mitigated by wholesale interbank trading markets. More broadly, mismeasured idiosyncratic *firm* disturbances could actually masquerade some unobserved/unmeasured borrower-side *network*-level risk. In the first stage of our analysis, it is possible that our measure of firm value-added shocks is in fact some kind of common local production network risk. This problem is mitigated considerably by the following three factors. First, when extracting firm value-added shocks, we impose a stringent combination of industry, location, and time fixed effects. Any unobserved network factor would have therefore to operate within the same year, industry, and county. Observations are sufficiently dispersed across geographical regions, with four out of 19 counties having more than 100,000 loan $\times$ year observations and the average being just over 50,000. In addition, our industry identifiers are very granular with only 3 out of the 99 two-digit industries having more than 100,000 loan $\times$ year observations. Second, we plotted in Figure 2.9 the pairwise cross-sectional correlation of firm shocks and found no indication of common shocks. Third, the presence of local production network effects would not invalidate our GIV approach. As long as the sequence of idisoyncratic shocks is unrelated to bank-side factors, our exogeneity assumption in equation (2.8) holds regardless of whether the firm-level shocks truly are firm-level or composite outcomes of very local network structures. The question of networks is therefore a matter of composition, not of identification.

2.6.2 Origins of Large Exposures

Credit concentration is an equilibrium object that is an outcome of more fundamental factors such as monitoring technology, risk preferences, information structure and expectations. Although writing down a micro-founded model of banking concentration is beyond the scope of this paper, discussing theoretical causes of concentration is useful for at least two reasons. First, the orthogonality of loan shares to idiosyncratic firm shocks is a key component of our empirical strategy, as highlighted in equation (2.8). Second, our paper is related to a large literature on the trade-off between risk concentration and economic performance and any normative implication would have to take into account the underlying causes of concentration.

One reason for the large degree of credit concentration observed in our data is home bias. Juelsrud and Wold (2020) document a substantial degree of within-county bias in the Norwegian banking system (see Figure 2.8). Using loan-level data, Juelsrud and Wold (2020) show that over 2003-2015 the average proportion of bank credit to firms that are headquartered in the same region as the lender was 55%. This compares to a random-assignment counterfactual of less than 10%, implying a home bias of 45%. Home bias is, of course, a perennial stylized fact in international finance, banking, and macroeconomics (Coeurdacier and Rey, 2013).

In what follows, we highlight three potential causes of portfolio concentration: asymmetric information, behavioral biases and the distribution of firm sizes. Asymmetric information and behavioral biases are also potential factors behind home bias in portfolios.

Asymmetric Information. Concentrated lending could be a by-product of persistent credit relationships. When information acquisition on new clients is costly, lenders may find it optimal to do business with a recurring set of borrowers, for instance by increasing the number of new commitments per relationship such as offering additional fixed-term loans or extending new credit lines (Sufi, 2007). Along the intensive margin, an increase in the exposure of an informed lender signals a higher quality of the underlying borrower, thereby reducing the cost of asymmetric information (Leland and Pyle, 1977). Van Nieuwerburgh and Veldkamp (2009) show in a rational inattention framework that investors may choose to learn only about assets for which they had an information advantage to start with (such as home assets), thus amplifying initial information asymmetries.

Ivashina (2009) proposes and tests a theory where the price of a loan is determined by a trade-off between diversification and asymmetric information. If a bank raises its exposure to a single borrower, two effects take place. First, the lender demands a higher premium for being more exposed to borrower-specific idiosyncratic risk. Second, assuming that expenditures on monitoring scale with exposure, concentration also reduces information asymmetry between the lender and the borrower, thus reducing the premium. In equilibrium, the price of the contract depends on the degree of information asymmetry and the magnitude of idiosyncratic fluctuations.

In our data, we observe a substantial degree of credit concentration. This is true at all levels of aggregation: single name, geographical, sectoral. In light of Ivashina (2009), this may suggest that Norwegian lenders attach large benefits to information acquisition. This is intuitive, given that the majority of firms in our data are not publically listed and are instead locally-focused small enterprises. Information collection and monitoring is therefore costly, and potentially increases with distance. This could explain both the regional home bias and the single name credit concentration facts. Banks' portfolio-level (over-)exposure to single borrowers solves the structural asymmetric information problem, but at the cost of an elevated vulnerability towards idiosyncratic shocks. Sensitivity to a given distribution of idiosyncratic borrower shocks (the loan share distribution $s_{i,j,t}$) can be explained by the banks' decision to invest in localized information acquisition, which is in turn driven by pre-determined factors such as the returns to information acquisition. This class of explanations validates our empirical approach.

Behavioral Biases. A second theory that could rationalize credit concentration rests on behavioral biases and overconfidence[30]. Huberman (2015) shows that some investors tend to ignore portfolio theory and invest in familar assets. In recent work, Bordalo et al. (2018) develop a theory of "diagnostic expectations" and apply it to canonical macroeconomic models of credit cycles. In their theory, agents persistently overweight future outcomes that have become more likely given recent data. Investors in this framework would make predictable forecast errors and could be "over-" or "underconfident" relative to the rational expectations benchmark.

The diagnostics expectations theory could potentially explain why lenders engage in excessive, unhedged credit concentration. Having established a credit relationship with a client that a bank trusts, the bank "specializes" in that client conditional on having exuberantly positive expectations about, for instance, the bank's own ability to pick superior portfolios and generate above-average returns based on own skill. Conditional on this expectations formation mechanism, the bank therefore "rationally" ignores the considerable downside risk of the strategy, i.e. the concentration risk. Our exogeneity assumption (2.8) is valid under diagnostics expectations at the bank-level. In that case, overexposure of bank i in firm j at time t is largely *independent* of the firm's present characteristics but is instead a function of i's subjective beliefs. Thus, behavioural biases of this kind would also be compatible with our empirical approach.

Distribution of firm sizes. Finally, credit concentration could be a by-product of the underlying firm size distribution also being fat tailed, which is definitely the case for Norway. Studies by Carvalho and Gabaix (2013) and Carvalho and Grassi (2019), among others, have shown that presence of a small number of large firms can explain a substantive percentage of aggregate macroeconomic fluctuations. Similarly, Gaubert and Itskhoki (2018) show that up to 20% of international export intensity can be attributed to granular firms. In the case of bank lending, if

[30] Fuster et al. (2010) review the extensive literature on the departures from rational expectations in finance and macroeconomics.

large firms are also large borrowers - a condition which is true in our data - the Pareto rate of the credit share distribution is driven by the Pareto rate of the firm size density. While this is a very natural explanation for the observed credit concentration and the one we pursue below in Section 2.7, it is worth nothing that in our data we also observe substantial heterogeneity in portfolio Herfindahl indices *across* banks, even among lenders of the *same* region. Banks do not all hold the same portfolio. Thus, (local) firm size concentration is not enough to completely explain either the home bias in bank lending or portfolio concentration. Financial frictions - be it informational, technological, or behavioral - are important as well.

For the purpose of the empirical analysis, we note that the three classes of models that we put forward to explain the origin of credit concentration (asymmetric information, behavioural biases, distribution of firm sizes) are all compatible with our empirical approach and identification strategy.[31]

2.7 Theoretical Motivation

Throughout the paper, we have exploited the stylized fact that the distribution of bank credit exhibits a fat tail. In this section, we provide one simple possible theoretical rationalization for this observation.[32] In the data, the right tail of the loan distribution is populated by a small number of very large loan contracts (as a share of the bank portfolio). These large loan contracts are almost always underwritten to big firms, a fact which we verify from our dataset. It is well known that the size distribution of firms is fat-tailed. If firm credit is a function of firm size, then we can precisely derive how the granularity of the firm distribution translates into the granularity of credit and affects portfolio-level outcomes.

A theoretical challenge encountered when formalizing this intuition is the fact that both firm loan and firm size distributions could potentially have infinite variances. In this particular case, standard central limit theorems break down. Following Gabaix (2011), we therefore resort to Lévy's generalized central limit theorems that can accommodate distributions with fat tails. In this section, we provide sufficient conditions for distributional parameter values to ensure that - assuming the firm size distribution has a fat tail - the firm credit distribution also has a fat tail.

A Simple Model of Firm Debt Suppose there are N firms in the economy[33]. Before production can begin, firms must obtain funding. By assumption, each firm *i* is cash-strapped and has to start the period by borrowing L_{it} from a bank. The growth rate of firm debt demand evolves according

[31] Those theories would have, however, different normative implications.

[32] As noted in section 2.6.2, other frictions would have to be added to fully account for the data.

[33] Alternatively, suppose there are N borrowers in a given bank's portfolio and we treat the bank as the "economy".

to:

$$\frac{\Delta L_{i,t+1}}{L_{it}} = \sigma_i \epsilon_{i,t+1} \tag{2.11}$$

where σ_i is the volatility of firm-level debt growth and $\epsilon_{i,t+1}$ are i.i.d. random variables. Economy-wide total stock of firm debt is:

$$D_t = \sum_i^N L_{it} \tag{2.12}$$

and growth of financial debt in the economy is

$$\frac{\Delta D_{t+1}}{D_t} = \sum_i^N \sigma_i \frac{L_{it}}{D_t} \epsilon_{i,t+1} \tag{2.13}$$

The variance of growth of total debt is the weighted sum of the variance of the volatility of idiosyncratic shocks to debt demand, with the shares equaling the squared share of firm *i*'s borrowing in the total economy. Assuming $\sigma_i = \sigma \; \forall$ i, we have:

$$\sigma_D = \left[\sum_i^N \sigma \left(\frac{L_{it}}{D_t} \right)^2 \right]^{\frac{1}{2}} \tag{2.14}$$

It is clear from equation 2.14 that the variance of total debt depends on the distribution of firm-level debt demand L_{it}. In our data, we see that firm-level borrowing is strongly positively correlated with firm size. Let firm size, proxied by either total assets or sales, be y_{it}. Assume idiosyncratic volatility of firm growth σ_y is constant and common to all firms. Following Gabaix (2011), we assume that $y_1, \ldots, y_N$ are drawn from a power law distribution:

$$\mathbb{P}(y > x) = (1 + x)^{-\alpha} \tag{2.15}$$

with the exponent $\alpha \geq 1$. Note that we set the location and scale parameters to zero and unity, for simplicity. In the literature, this precise specification of a power law corresponds to a Pareto distribution of Type II.

Now, we assume a specific functional form for the amount of borrowing L_{it} as a function of size y_{it}:

$$L_{it} = y_{it}^{\lambda_i} \tag{2.16}$$

where $\lambda_i > 0 \, \forall$i. We proceed with the assumption that $\lambda_i = \lambda$ is homogenous across all firms.

Drawing from the literature on statistics, economics, and actuarial sciences, we know that once y_i follows a power law, then y_i^{λ} follows a Champernowne (1952) distribution, also known as the Burr Type XII, with parameters $\{\tau, \alpha\}$ where $\tau = 1/\lambda$ (Rodriguez, 1976). In economics, this distribution is commonly referred to as the Singh-Maddala (SM) density (Singh and Maddala, 1976). It has been used widely to model household income and wealth inequality. Formally:

$$\mathbb{P}(\mathrm{L} > \mathrm{x}) \sim (1 + \mathrm{x}^{\tau})^{-\alpha} \tag{2.17}$$

with $\tau > 0$. For the special case of $\tau = 1$, firm debt becomes linear in size, the distribution collapses to a Pareto Type II, and we are back to Gabaix (2011). In general, the rate of decay of σ_{D}, as the sample size grows, will depend on the value of structural parameters. For the special case of $1 < \tau\alpha < 2$, the SM random variable has an infinite variance and standard limit theorems break down. There is therefore a direct link between the fat tail of the firm distribution and of the credit distribution. This result is summarized in our main proposition below:

Proposition 3. *Let firm sizes* $\mathrm{y}_{\mathrm{i}} \ldots \mathrm{y}_{\mathrm{N}}$ *be drawn from a power law distribution with exponent* $\alpha \geq 1$. *Suppose each firm has non-rationed access to the credit market, through which on demand it borrows a fraction* $\mathrm{y}^{\lambda-1}$ *of its size, with* $\lambda>0$. *Firm-level borrowing is thus* $\mathrm{L} = \mathrm{y}^{\lambda}$, *which grows with a constant idiosyncratic volatility* σ. L *follows the Singh-Maddala distribution with power and shape parameters* $\{\alpha, \tau\}$*:*

$$\mathbb{P}(\mathrm{L} > \mathrm{x}) \sim (1 + \mathrm{x}^{\tau})^{-\alpha}$$

with $\tau = 1/\lambda$. *Then, as* $\mathrm{N} \to \infty$*:*

- *For* $1 < \alpha\tau < 2$, *by the Lévy's central limit theorem, the volatility of aggregate debt D is given by* $\sigma_{\mathrm{D}} \sim \sigma \frac{1}{\mathrm{N}^{1-1/(\alpha\tau)}} \sqrt{\eta}$, *where* η *is a Lévy random variable with exponent* $\frac{\alpha\tau}{2}$
- *For* $\alpha\tau \geq 2$, *by the Lindeberg-Lévy classical central limit theorem, the volatility of aggregate debt D is given by* $\sigma_{\mathrm{D}} \sim \sigma \frac{1}{\mathrm{N}^{1/2}} \sqrt{\eta}$, *where* η *is a constant*

Proof. The strategy of the proof follows closely Appendix 1 and Proposition 2 in Gabaix (2011). First, we show that L, which follows the Singh-Maddala distribution, satisfies Assumptions 1-2 below:

Assumption 1: $\lim_{\mathrm{l}\to\infty} \mathbb{P}(\mathrm{L}_1 > \mathrm{x})/\mathbb{P}(|\mathrm{L}_1| > \mathrm{x}) = \kappa \in [0, 1]$

Assumption 2: $\mathbb{P}(|\mathrm{L}_1| > \mathrm{x}) = \mathrm{x}^{-\alpha}\mathrm{B}(\mathrm{x})$ with $\mathrm{B}(\mathrm{x})$ a slow-moving function.

Assumption 1 is verified trivially because SM is defined on the non-negative real line. Assumption 2 holds once we re-write: $\mathbb{P}(|\mathrm{L}_1| > \mathrm{x}) = \mathrm{x}^{-\alpha}(\frac{\mathrm{x}}{1+\mathrm{x}^{\tau}})^{\alpha}$. So, $\mathrm{B}(\mathrm{x}) = (\frac{\mathrm{x}}{1+\mathrm{x}^{\tau}})^{\alpha}$. For $\tau = 1$, the function is clearly slow-moving. Generally, for $\tau > 0$ we must show that:

$$\lim_{\mathrm{x}\to\infty} \mathrm{B}(\mathrm{tx})/\mathrm{B}(\mathrm{x}) = \frac{\lim \mathrm{B}(\mathrm{tx})}{\lim \mathrm{B}(\mathrm{x})} = 1 \tag{2.18}$$

for any t>0 and for as long as the denominator is $\neq 0$. $\lim_{x\to\infty} B(x) = \lim_{x\to\infty}\left[\frac{x}{1+x^{\tau}}\right]^{\alpha} = \lim_{x\to\infty}\left[\frac{1}{1/x+x^{\tau-1}}\right]^{\alpha} = 1$. Similarly for B(tx).

Next, we construct three sequences (a_n, b_n, s_n) that constitute the infinite sum across firms. $a_n = \inf(x : \mathbb{P}(|L_1| > x) \leq 1/N) \sim \left(N^{1/\alpha} - 1\right)^{1/\tau} \approx N^{\frac{1}{\alpha\tau}}$. $b_n = n\mathbb{E}\left(L_1 1_{|L|\leq a_n}\right) = 0$. And $s_n = \sum_i^N L_i$. Thus:

$$\lim_{N\longrightarrow\infty}\left(N^{\frac{1}{\alpha\tau}}\right)^{-1}\sum_i^N L_i \xrightarrow{d} \eta \sim \text{Lévy}(\alpha\tau) \tag{2.19}$$

In the remainder of the proof, we apply equation (2.19) to the case of constant σ, i.e. when firm-liability volatility is constant over time and not correlated cross-sectionally. When $\alpha\tau > 2$, standard Lindeberg-Lévy applies. When $1 < \alpha\tau < 2$, the loan portfolio Herfindahl decays according to:

$$N^{1-\frac{1}{\alpha\tau}}\frac{\left(N^{\frac{-2}{\alpha\tau}}\sum_i^N L_i^2\right)^{1/2}}{N^{-1}\sum_i^N L_i} \xrightarrow{d} \frac{\text{Lévy}^{1/2}}{\mathbb{E}(L)} \tag{2.20}$$

When $1 < \alpha\tau < 2$, the denominator (mean of Singh-Maddala) is finite. Since firm-level volatilities are constant, and Lévy is a stable random variable, the volatility of loan growth will be therefore decaying at the rate proportional to $N^{1-\frac{1}{\alpha\tau}}$:

$$\sigma_D \sim \frac{1}{N^{1-1/(\alpha\tau)}}\text{Lévy}^{1/2}\sigma \tag{2.21}$$

For $\tau = 1$ we are in the special case of Singh-Maddala collapsing to the Pareto II distribution and standard results in Gabaix (2011) are obtained up to the slow-moving function B(.). For $\tau \neq 1$ but $\tau > 0$, the sufficient statistic for the comparison of rates of convergence across finite and infinite variance cases is $\alpha\tau$.

□

Our notation means that $\sigma_D \sim \sigma\frac{1}{N^{1-1/(\alpha\tau)}}\sqrt{\eta}$ implies convergence in distribution of $\sigma_D N^{1-1/(\alpha\tau)}$ to $\sigma\sqrt{\eta}$, where η is a stable Lévy random variable. What we have shown is that the distribution of firm debt could have either thin or fat tails. If $\alpha\tau \geq 2$, σ_D decays according to $\frac{1}{\sqrt{N}}$. However, if $1 < \alpha\tau < 2$, then σ_D decays at the rate of $\frac{1}{N^{1-\frac{1}{\alpha\tau}}}$, i.e. more slowly. In this case, idiosyncratic shocks to borrowers could drive the total debt portfolio and, as in our main empirical experiments, affect aggregate outcomes.

2.7.1 Model Parameter Estimation

In this section, we test whether the parameter restriction $1 < \alpha\tau < 2$ can be supported by our data. First, we fit the Generalized Pareto density into the size distribution of firms. Most studies in the literature treat sales as the size proxy. We, apart from sales, also consider total equity and total assets as alternative size proxies that could be relevant for deciding on how much bank credit to request. This step grants us three estimates of α. Second, we back out firm-specific λ_i directly from equation (2.16) and then take the median of the resulting distribution. We conduct this step for all three definitions of size as well. As a result, we have three estimates for $\alpha\tau$ - the sufficient statistic that determines the speed of decay of σ_D.

Table 2.26 in the Online Appendix reports the results from maximum likelihood estimation of α and other parameters. Our estimates confirm that the $1 < \alpha\tau < 2$ restriction is supported in the data. We find that α is in the [1.26, 1.49] range and $\alpha\tau$ is between 1.38 and 1.64, i.e. firmly within the (1,2) bounds. Our estimation results suggest that both the firm size and the firm loan distributions can be reasonably approximated with fat-tailed densities. The aggregate credit distribution can be affected by firm-level disturbances: credit risk is granular.

2.8 Conclusion

This paper has developed the first, bottom-up causal quantification of single-name credit concentration risk. To the best of our knowledge, we are the first to provide evidence on *single-name* (individual firm-level) as opposed to sectoral or geographic exposure risk. Empirically, we show there is a causal link between idiosyncratic firm shocks and returns on bank credit. Unexpected shocks to firm value-added affect loan-level and bank-level performance. We capture strong asymmetries associated with the debt contract structure by showing that negative firm shocks lead to a reduction in bank returns, while positive shocks have zero impact. We explored numerous dimensions of heterogeneity at all levels of aggregation.

We find strong evidence of a second-level pass-through effect of granular borrower risk onto other firms. Banks, in response to negative shocks to their granular borrowers, cut credit supply and increase interest rates on loans to their non-granular borrowers. Affected non-granular firms, in turn, reduce investment in physical capital. Affected firms are also more likely to file for bankruptcy following a negative granular shock to their credit provider. These results suggest that single-name credit concentration risk carries significant implications for the macroeconomy.

The first key message of the paper is therefore that idiosyncratic firm shocks do not wash out and still matter at the level of the bank portfolio. Conventional wisdom that banks are subject only to aggregate risk due to pooling and the law of large number is not borne out in the data.

Concentration risk matters quantitatively. Our evidence from non-interest income data further suggests that banks do not compensate for loan book losses through earnings from alternative sources such as derivatives or equity holdings. The second key message of the paper is that there are important *granular credit risk spillovers* affecting the real economy.

Methodologically, we make progress on identification of firm demand-side shocks at the level of bank portfolios by employing the "granular instrument variable" approach developed in recent work by Gabaix and Koijen (2020a,b). This method takes advantage of the fact that the distribution of loan shares features a fat tail and allows for a tightly-identified pass-through of granular risk.

Our results have implications for the regulation of large credit exposures. Our pass-through estimates in Table 2.3 could be used to compute the *granular value-at-risk*, i.e. the bank capital that is at risk if a granular borrower suffers a bad negative shock. Our estimate of the loan share Pareto power of section 2.7 could be used as a tool for understanding when banks are becoming prone to granular credit risk. A drop in the Pareto power estimate to 2 or below could constitute a regulatory "red flag". In practice, the parameter could be computed for each financial institution in the cross-section. The system wide weighted average Pareto estimate could become a novel time-series indicator of aggregate concentration whose changes could track fluctuations in *system wide credit concentration risk*.

References

Acharya, V., I. Hasan, and A. Saunders (2006): "Should banks be diversified? Evidence from individual bank loan portfolios," *Journal of Business*, 79.

Agarwal, S., R. Correa, B. Morais, J. Roldan, and C. Ruiz-Ortega (2020): "Owe a Bank Millions, the Bank Has a Problem: Credit Concentration in Bad Times," *International Finance Discussion Papers*, 1288.

Amiti, M., O. Itskhoki, and J. Konings (2019): "International Shocks, Variable Markups, and Domestic Prices," *Review of Economic Studies*.

Amiti, M. and D. Weinstein (2018): "How Much Do Idiosyncratic Bank Shocks Affect Investment? Evidence from Matched Bank-Firm Data," *Journal of Political Economy*, Forthcoming.

Beck, T., O. D. Jonghe, and K. Mulier (2017): "Bank sectoral concentration and (systemic) risk: Evidence from a worldwide sample of banks," *CEPR Working Paper*, DP12009.

BIS (2013): "Supervisory framework for measuring and controlling large exposures," *Consultative Document*.

Bordalo, P., N. Gennaioli, and A. Shleifer (2018): "Diagnostic Expectations and Credit Cycles," *The Journal of Finance*, 73(1), 199–227.

Boyd, J. H. and E. C. Prescott (1986): "Financial intermediary-coalitions," *Journal of Economic Theory*, 38.

Bremus, F., C. Buch, K. Russ, and M. Schnitzer (2018): "Big Banks and Macroeconomic Outcomes: Theory and Cross-Country Evidence of Granularity," *Journal of Money, Credit and Banking*, Forthcoming.

Bruno, V. and H. Shin (2015): "Capital flows and the risk-taking channel of monetary policy," *Journal of Monetary Economics*, 71.

Carvalho, V. and X. Gabaix (2013): "The Great DiversiÖcation and its Undoing," *American Economic Review*, 103(5).

Carvalho, V. M. and B. Grassi (2019): "Large Firm Dynamics and the Business Cycle," *American Economic Review*, 109(4), 1375–1425.

Champernowne, D. G. (1952): "The Graduation of Income Distributions," *Econometrica*, 20, 591–615.

Chang, B., M. Gomez, and H. Hong (2021): "Sorting Out the Real Effects of Credit Supply," *NBER Working Paper*, 28842.

Chetty, R., A. Looney, and K. Kroft (2009): "Salience and taxation: Theory and evidence," *American economic review*, 99, 1145–77.

Chodorow-Reich, G. (2014): "The Employment Effects of Credit Market Disruptions: Firm-level Evidence from the 2008-09 Financial Crisis," *Quarterly Journal of Economics*, 129.

Choi, N., M. Fedenia, H. Skiba, and T. Sokolyk (2017): "Portfolio concentration and performance of institutional investors worldwide," *Journal of Financial Economics*, 123.

Coeurdacier, N. and H. Rey (2013): "Home Bias in Open Economy Financial Macroeconomics," *Journal of Economic Literature*, 51(1).

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

Davila, E. and A. Walther (2020): "Does Size Matter? Bailouts with Large and Small Banks," *Journal of Financial Economics*, 136(1).

Dewachter, H., J. Tielens, and J. V. Hove (2020): "The Production Network Sources of Bank Influence," *Working Paper*.

di Giovanni, J., A. Levchenko, and A. Mejean (2014): "Firms, destinations, and aggregate fluctuations," *Econometrica*, 82(4).

di Giovanni, J., A. Levchenko, and I. Mejean (2018): "The Micro Origins of International Business Cycle Comovement," *American Economic Review*, 108(1).

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

Elliott, M., C.-P. Georg, and J. Hazell (2020): "Systemic Risk Shifting in Financial Networks," *Journal of Economic Theory (Forthcoming)*.

FAGERENG, A., L. GUISO, AND L. PISTAFERRI (2018): "Portfolio Choices, Firm Shocks, and Uninsurable Wage Risk," *Review of Economic Studies*, 85(1).

FOSTER, L., J. HALTIWANGER, AND C. SYVERSON (2008): "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review*, 98(1).

FOSTER, L. S., C. A. GRIM, J. HALTIWANGER, AND Z. WOLF (2017): "Macro and Micro Dynamics of Productivity: From Devilish Details to Insights," .

FUSTER, A., D. LAIBSON, AND B. MENDEL (2010): "Natural Expectations and Macroeconomic Fluctuations," *Journal of Economic Perspectives*, 24(4).

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

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

GABAIX, X. AND R. KOIJEN (2020a): "Granular Instrumental Variables," *Manuscript*.

——— (2020b): "In search of the origins of financial fluctuations: The inelastic markets hypothesis," *Manuscript*.

GAUBERT, C. AND O. ITSKHOKI (2018): "Granular Comparative Advantage," *NBER Working Paper*.

GOETZ, M. R., L. LAEVEN, AND R. LEVINE (2016): "Does the Geographic Expansion of Banks Reduce Risk?" *Journal of Financial Economics*, 120(2).

GREENWALD, D., J. KRAINER, AND P. PAUL (2020): "The Credit Line Channel," *Working Paper*.

GUISO, L., L. PISTAFERRI, AND F. SCHIVARDI (2005): "Insurance within the Firm," *Journal of Political Economy*, 113(5).

HSIEH, C. AND P. KLENOW (2009): "Misallocation and Manufacturing TFP in China and India," *The Quarterly Journal of Economics*, 124(4).

HUBERMAN, G. (2015): "Familiarity Breeds Investment," *The Review of Financial Studies*, 14, 659–680.

HUREMOVIC, K., G. JIMENEZ, E. MORAL-BENITO, J. L. PEYDRO, AND F. VEGA-REDONDO (2020): "Production and Financial Networks in Interplay: Crisis Evidence from Supplier-Customer and Credit Registers," *Working Paper*.

IVASHINA, V. (2009): "Asymmetric information effects on loan spreads," *Journal of Financial Economics*, 92(2).

JAMILOV, R. (2021): "A Macroeconomic Model with Heterogeneous Banks," *Manuscript*.

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

KHWAJA, A. I. AND A. MIAN (2008): "Tracing the impact of bank liquidity shocks: Evidence from an emerging market," *American Economic Review*, 98, 1413–42.

LEARY, M. AND M. ROBERTS (2014): "Do Peer Firms Affect Corporate Financial Policy?" *The Journal of Finance*, 69(1).

LELAND, E. AND D. H. PYLE (1977): "Information asymmetries, financial structure, and financial intermediation," *Journal of Finance*, 32.

LIN, H. AND D. PARAVISINI (2012): "The Effect of Financing Constraints on Risk," *Review of Finance*.

MENDICINO, C., K. NIKOLOV, J. RUBIO-RAMIREZ, J. SUAREZ, AND D. SUPERA (2019): "Twin Default Crises," *Working Paper*.

PARAVISINI, D., V. RAPPOPORT, AND P. SCHNABL (2020): "Specialization in Bank Lending: Evidence from Exporting Firms," *Working Paper*.

PEEK, J. AND E. S. ROSENGREN (2000): "Collateral Damage: Effects of the Japanese Bank Crisis on Real Activity in the United States," *American Economic Review*, 90.

RODRIGUEZ, R. (1976): "A Guide to the Burr Type XII Distributions," *Institute of Statistics Mimeo Series 1064*.

SHARPE, S. (1990): "Asymmetric Information, Bank Lending and Implicit Contracts: A Stylized Model of Customer Relationships," *Quarterly Journal of Economics*, 45.

SINGH, S. AND G. MADDALA (1976): "A Function for Size Distribution of Incomes," *Econometrica*, 44(5).

STAVRAKEVA, V. (2019): "Optimal Bank Regulation and Fiscal Capacity," *Review of Economic Studies*, Forthcoming.

STOCK, J. AND M. YOGO (2005): "Testing for Weak Instruments in Linear IV Regression," 80–108.

SUFI, A. (2007): "Bank Lines of Credit in Corporate Finance: An Empirical Analysis," *Review of Financial Studies*, 22.

VAN NIEUWERBURGH, S. AND L. VELDKAMP (2009): "Information immobility and the home bias puzzle," *The Journal of Finance*, 64, 1187–1215.

WILLIAMSON, S. (1987): "Costly Monitoring, Loan Contracts, and Equilibrium Credit Rationing," *Quarterly Journal of Economics*, 102.

Appendix

2.9 Factor Analysis

2.9.1 Factor Extraction at the Firm Level

Our baseline firm shock measure is the residual $\epsilon_{j,t}$ obtained from estimating equation (2.1) in main text, repeated here:

$$\ln VA_{j,t} = \alpha_{j,t,s(j),z(j)} + \beta_1 \ln K_{j,t} + \beta_2 \ln W_{j,t} + \lambda' X_{j,t} + \epsilon_{j,t}. \tag{2.22}$$

The residual $\epsilon_{j,t}$, although orthogonal to a range of time-varying firm characteristics and fixed effects, may still contain components which are common across firms. To address this concern we now consider a robustness exercise in which we extract both parametric and non-parametric factors explicitly. Formally, we assume that the residual can be expressed as:

$$\epsilon_{j,t} = \delta^{x}_{j,t}{}' \eta^{x}_{t} + \delta'_{j} \eta_{t} + u_{j,t}, \tag{2.23}$$

for a vector of parametric η^{x}_{t} and non-parametric η_t factors. For the parametric factors, the firm-specific time-varying loading vector $\delta^{x}_{j,t}$ is assumed to be a function of observable firm characteristics. For the non-parametric factors we assume a constant firm-specific loading vector δ_j. The goal is to estimate both common components ($\delta^{x}_{j,t}{}' \eta^{x}_{t}$ and $\delta'_{j} \eta_{t}$) and to replace our firm shock measure $\epsilon_{j,t}$ with a more robust alternative $u_{j,t}$.

We proceed in two steps. First, we extract parametric common components by estimating a richer version of equation (2.22), in which we interact all time-varying firm-specific regressors ($\ln K_{j,t}$, $\ln W_{j,t}$, $X_{j,t}$) with year dummies. Hence, $\delta^{x}_{j,t}$ is given by the vector of explanatory variables in equation (2.22). Formally, we re-estimate equation (2.22) assuming time-varying coefficients:[34]

$$\ln VA_{j,t} = \alpha_{j,t,s(j),z(j)} + \beta_{1,t} \ln K_{j,t} + \beta_{2,t} \ln W_{j,t} + \lambda'_{t} X_{j,t} + \check{\epsilon}_{j,t}. \tag{2.24}$$

In the second step, we perform Principal Component Analysis (PCA) on the residual $\check{\epsilon}_{j,t}$ by estimating:

$$\check{\epsilon}_{j,t} = \delta'_{j} \eta_{t} + u_{j,t} \tag{2.25}$$

Since our firm panel is unbalanced, we employ an iterative Expectation Maximization (EM) algorithm as in Gabaix and Koijen (2020), and estimate principal components recursively. Starting

[34] We make one adjustment relative to the specification in equation (2.22), by replacing the quadratic age specification with one-year age fixed effects.

Table 2.9: **Loan Outcomes with Firm Factors Extraction**

	(1)	(2)	(3)
	Dep. Var.: Return on Loan		
(1) Firm Shock: $\check{\epsilon}_{j,t}$	0.307	0.307	0.333
	(0.016)	(0.017)	(0.018)
(2) Firm Shock: $u^1_{j,t}$	0.279	0.279	0.299
	(0.016)	(0.017)	(0.018)
(3) Firm Shock: $u^2_{j,t}$	0.239	0.237	0.255
	(0.016)	(0.017)	(0.018)
Bank x Industry x Year FE	No	Yes	No
Bank x Industry x Year x Loan-type x County FE	No	No	Yes

Notes: This table reports results from the regression of loan-level returns on loans on three alternative measures of idiosyncratic firm shocks. Row (1) refers to the shock measure after extracting parametric common components. Row (2) refers to the shock measure after extracting parametric common components and one latent common component. Row (3) refers to the shock measure after extracting parametric common components and two latent common components. All shocks have been normalized by their standard deviations. Standard errors (in parentheses) are double clustered at the firm-year level.

with the first factor, the algorithm repeatedly regresses $\check{\epsilon}_{j,t}$ on η^1_t and then $\check{\epsilon}_{j,t}$ on δ^1_j until convergence. For factors $f = 2, \ldots, f^{max}$, least squares iterations are performed on the remaining residual from equation (2.25) after extracting $f-1$ components, denoted $u^{f-1}_{j,t}$.[35] In our analysis below we consider $f^{max} = 2$ components and denote by $u^1_{j,t}$ and $u^2_{j,t}$ the residuals obtained after extracting one and two factors, respectively.[36]

We then run our loan-level regressions based on equation (2.2) in main text with the three new estimated firm shock measures: $\check{\epsilon}_{j,t}$, $u^1_{j,t}$ and $u^2_{j,t}$. In other words, we substitute the baseline shock variable $\epsilon_{j,t}$ with potentially more refined and idiosyncratic versions. In order to obtain bank-level estimates, we proceed as in the main text. First, we aggregate by computing loan size-weighted averages of the three new shock measures $\bar{\check{\epsilon}}_{i,t}$, $\bar{u}^1_{i,t}$, and $\bar{u}^2_{i,t}$. Second, we construct three new Granular IVs using equation (2.6). Third, we run our two-stage panel regressions for $\bar{\check{\epsilon}}_{i,t}$, $\bar{u}^1_{i,t}$, and $\bar{u}^2_{i,t}$, instrumenting each with their respective $GIV^{\check{\epsilon}}_{i,t}$, $GIV^{u^1}_{i,t}$, and $GIV^{u^2}_{i,t}$.

Table 2.9 reports loan outcomes after factor extraction. Columns (1)-(3) are based on the same set of controls and fixed effects as columns (3)-(5) in Table 2.2. Rows (1)-(3) show results for the three new shock measures. Recall that baseline estimates from Table 2.2 are in the 0.334-0.361

[35]Following the suggestion in Stock and Watson (2016), iterations are initiated with factors that are extracted from the balanced sub-sample of firms.

[36]The f^{max} threshold is chosen by performing a standard PCA on a balanced sub-sample of firms, and applying the IC_{p2} criterion in Bai and Ng (2002) to determine the number of factors.

range. We see that after the extraction of parametric and two non-parametric factors, estimates are still large, statistically significant, and quantitatively very close to our baseline results.

Table 2.10 reports results at the bank level. Columns (1)-(8) are based on the same specifications and sets of controls and fixed effects as columns (1)-(8) in Table 2.3 from main text. Recall that baseline estimates from Table 2.3 are roughly 0.117 and 0.180 for the specifications with pooled and only negative shocks, respectively. We find that our strictest model, which extracts parametric and two non-parameteric factors, leads to estimates of 0.083 and 0.119 for pooled and only negative shocks specifications, respectively. All coefficients are very similar to our baseline results and are statistically significant at least at the 5% level.

We now consider an alternative approach where instead of replacing the baseline shock measure $\epsilon_{j,t}$ itself, we keep $\epsilon_{j,t}$ as the shock variable and build the Granular IV based on the three new shocks $GIV^{\check{\epsilon}}_{i,t}$, $GIV^{u^1}_{i,t}$, and $GIV^{u^2}_{i,t}$. In other words, we keep the same endogenous regressor but instrument it with new, more robust instruments. Results are reported in Table 2.11. All estimates are quantitatively in line with our baseline results. Coefficients from specifications with pooled or negative only shocks are all statistically significant at least at the 5% level.

2.9.2 Factor Extraction at the Bank Level

By subtracting the unweighted mean from bank-level weighted firm shocks, our Granular IV in equation (2.6) of the main text removes a common bank factor with loadings δ_i that are assumed to be identical across firms. In this section, we relax the assumption of homogeneous loadings and consider a generalized GIV by extracting common factors at the bank level. Formally, we now extract common components for each bank separately. This implies running the EMPCA algorithm separately on each bank's sample of borrowers, i.e. for all firms j borrowing from bank i at time t:

$$\check{\epsilon}^{d}_{i,j,t} = \delta'_{i,j}\eta_{i,t} + u_{i,j,t}\,,\ \forall i \tag{2.26}$$

where $\check{\epsilon}^{d}_{i,j,t}$ denotes the demeaned firm shock $\check{\epsilon}_{j,t}$ (the residual net of parametric factors from equation (2.24)). The demeaning is performed cross-sectionally at the bank level, such that:

$$\check{\epsilon}^{d}_{i,j,t} = \check{\epsilon}_{j,t} - \frac{1}{N_{i,t}}\sum_{j}\check{\epsilon}_{j,t}$$

Table 2.10: **Bank Outcomes with Firm Factors Extraction - New Shocks, New GIV**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable: Bank Return on Loans (RoA)							
	OLS		Instrumented with GIV					
	Pooled	Pooled	Pooled	Positive	Negative	Pooled	Positive	Negative
(1) Granular Credit Shock: $\breve{\bar{\epsilon}}_{j,t}$	0.118	0.125	0.106	0.015	0.212	0.105	0.027	0.186
	(0.027)	(0.026)	(0.035)	(0.081)	(0.075)	(0.030)	(0.071)	(0.073)
(2) Granular Credit Shock: $\bar{u}^1_{j,t}$	0.092	0.092	0.079	-0.117	0.160	0.072	-0.087	0.136
	(0.025)	(0.024)	(0.031)	(0.078)	(0.073)	(0.029)	(0.075)	(0.068)
(3) Granular Credit Shock: $\bar{u}^2_{j,t}$	0.106	0.100	0.090	-0.082	0.136	0.083	-0.067	0.119
	(0.027)	(0.025)	(0.032)	(0.072)	(0.058)	(0.029)	(0.072)	(0.053)
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bank Controls	No	Yes	No	No	No	Yes	Yes	Yes

Notes: This table reports results from regressing bank-level return on loans on bank-level aggregated firm shocks. Columns (1)-(2) are standard OLS, while columns (3)-(8) instrument the weighted shock with a granular IV. Row (1) is based on the shock $\breve{\epsilon}_{j,t}$ and instrument $GIV^{\breve{\epsilon}}_{i,t}$, which refer to the residual after extracting parametric common components. Row (2) is based on the shock $u^1_{j,t}$ and instrument $GIV^{u^1}_{i,t}$, which refer to the residual after extracting parametric common components and one latent common component. Row (3) is based on the shock $u^2_{j,t}$ and instrument $GIV^{u^2}_{i,t}$, which refer to the residual after extracting parametric common components and two latent common components. Standard errors (in parentheses) are clustered at the bank level.

where $N_{i,t}$ denotes bank i′s number of corporate borrowers j in year t.[37] We extract up to $f = 2$ factors, following the algorithm outlined in 2.9.1, and keep the residuals $u^f_{j,t}$, with $f \in \{1, 2\}$.[38]

Our main exercise is to use the extracted bank factors $\eta^1_{i,t}$ and $\eta^2_{i,t}$ as explicit controls in our bank-level regressions. This approach is similar to the application that is proposed in Gabaix and Koijen (2020). Specifically, we run the following specification:

$$R^b_{i,t} = \alpha_i + \alpha_t + \beta_1 \hat{u}_{i,t} + \beta_2 \eta^1_{i,t} + \beta_3 \eta^2_{i,t} + \nu_{it} \tag{2.27}$$

[37]Notice that demeaning the firm shock prior to constructing the loan-size weighted shock is identical to constructing the GIV as the difference between the size-weighted minus unweighted raw firm shocks ϵ. Consequently, if we extract zero factors in equation (2.26) we get the same bank-level estimates as in row (1) of Table 2.10. Hence $\sum_j s_{i,j,t}\breve{\epsilon}^d_{i,j,t}$ is identical to the $GIV_{i,t}$ based on $\breve{\epsilon}_{j,t}$.

[38]Because very few banks in our sample have fully balanced sub-samples (portfolios) with many borrowers, we now initiate the algorithm with random guesses of realizations for each factor f ($\eta^f_1, \eta^f_2, ..., \eta^f_T$) with 100 different seeds and pick the specification that produces the lowest average sum of squared residuals $u^{f^{max}}_{j,t}$ after extracting $f^{max} = 2$ components.

Table 2.11: **Bank Outcomes with Firm Factors Extraction - Old Shocks, New GIV**

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Variable: Bank Return on Loans (RoA)					
	Pooled	Positive	Negative	Pooled	Positive	Negative
(1) $GIV^{\check{\epsilon}}_{i,t}$	0.110	0.003	0.182	0.111	0.035	0.165
	(0.035)	(0.078)	(0.071)	(0.030)	(0.070)	(0.068)
(2) $GIV^{u^1}_{i,t}$	0.114	-0.021	0.216	0.112	0.035	0.189
	(0.032)	(0.092)	(0.074)	(0.028)	(0.095)	(0.065)
(3) $GIV^{u^2}_{i,t}$	0.144	0.039	0.266	0.133	0.061	0.234
	(0.038)	(0.140)	(0.084)	(0.032)	(0.135)	(0.071)
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Bank Controls	No	No	No	Yes	Yes	Yes
Instrumented with GIV	Yes	Yes	Yes	Yes	Yes	Yes

Notes: This table reports results from regressing bank-level return on loans on bank-level aggregated firm shocks $\bar{\epsilon}_{i,t}$ instrumented by three alternative Granular IVs. In row (1) the GIV is based on $\bar{\check{\epsilon}}_{i,t}$, which refers to the shock measure after extracting parametric common components. In row (2) the GIV is based on $\bar{u}^1_{i,t}$, which refers to the shock measure after extracting parametric common components and one latent common component. In row (3) the GIV is based on $\bar{u}^1_{i,t}$, which refers to the shock measure after extracting parametric common components and two latent common components. Standard errors (in parentheses) are clustered at the bank level.

Results are reported in Table 2.12. In columns (2) and (6)-(8) we have added the two extracted factors to the list of our usual bank-level controls. Results are essentially unchanged relative to our baseline estimation. This indicates that endogeneity issues due to unobserved time-varying bank factors are minor.

As an additional and final robustness check, we focus on the residuals extracted from equation (2.26). Similarly to our aggregate factor extraction exercise in Section 2.9.1, we run two separate specifications. First, we substitute the baseline endogenous covariate $\bar{\check{\epsilon}}_{j,t}$ with the two new measures $\bar{u}^1_{i,t}$ and $\bar{u}^2_{i,t}$, which are robust to heterogeneous $\delta_{i,j}$. For these two shock measures, we construct new Granular IVs the usual way: $GIV^{u^1}_{i,t}$ and $GIV^{u^2}_{i,t}$. Second, instead of replacing the baseline shock measure $\bar{\check{\epsilon}}_{j,t}$, we retain it as the shock variable but instrument it with the $GIV^{u^1}_{i,t}$ or $GIV^{u^2}_{i,t}$. Results from the two exercises are reported in Tables 2.13 and 2.14. Our main focus is on columns (6) and (8) in both tables. We see that all results remain qualitatively robust, however the point estimates drop slightly and the negative-only estimates become noisier.[39]

[39]Because the panel is highly unbalanced, the effective time dimension is very small. This means that if use more than two factors, we may be over-fitting the data. In other words, with more factors we could be falsely re-labeling truly

Table 2.12: **Bank Factors Extraction - Controlling for Factors Directly**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable: Bank Return on Loans (RoA)							
	OLS		Instrumented with GIV					
	Pooled	Pooled	Pooled	Positive	Negative	Pooled	Positive	Negative
Granular Credit Shock: $\bar{\check{\epsilon}}_{j,t}$	0.125	0.123	0.105	0.027	0.186	0.104	0.024	0.181
	(0.026)	(0.025)	(0.030)	(0.071)	(0.073)	(0.029)	(0.071)	(0.073)
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bank Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
$\eta_{i,t}$ controls	No	Yes	No	No	No	Yes	Yes	Yes

Notes: This table reports the results from regressing bank-level return on loans on bank-level aggregated firm shocks. The firm level shock is based on $\check{\epsilon}_{j,t}$. Column (1) and (3)-(5) repeats the estimation with bank controls from Table 2.10. The other columns adds the first two latent bank-level factors obtained from running PCA separately on each banks' sample of borrowers using equation (2.26) to the set of bank controls.

idiosyncratic variation as common shocks, which in turn makes estimation less accurate. Gabaix and Koijen (2020) discuss a similar issue.

Table 2.13: **Bank Factors Extraction - New Shocks, New GIV**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable: Bank Return on Loans (RoA)							
	OLS		Instrumented with GIV					
	Pooled	Pooled	Pooled	Positive	Negative	Pooled	Positive	Negative
(1) Granular Credit Shock: $\bar{u}^1_{i,t}$	0.050	0.048	0.065	0.092	0.143	0.061	0.110	0.131
	(0.025)	(0.022)	(0.025)	(0.067)	(0.044)	(0.021)	(0.051)	(0.038)
(2) Granular Credit Shock: $\bar{u}^2_{i,t}$	0.035	0.026	0.056	-0.024	0.123	0.045	-0.025	0.105
	(0.022)	(0.020)	(0.024)	(0.060)	(0.057)	(0.020)	(0.053)	(0.053)
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bank Controls	No	Yes	No	No	No	Yes	Yes	Yes

Notes: This table reports results from regressing bank-level return on loans on bank-level aggregated firm shocks. Columns (1)-(2) are standard OLS, while columns (3)-(8) instrument the weighted shock with a granular IV. Rows (1)-(2) are based on firm shocks $u^1_{i,j,t}$ and $u^2_{i,j,t}$ obtained from running PCA separately on each bank's sample of borrowers using equation (2.26). These are the residuals remaining after extracting 1 and 2 common components, respectively, at the bank level. Standard errors (in parentheses) are clustered at the bank level.

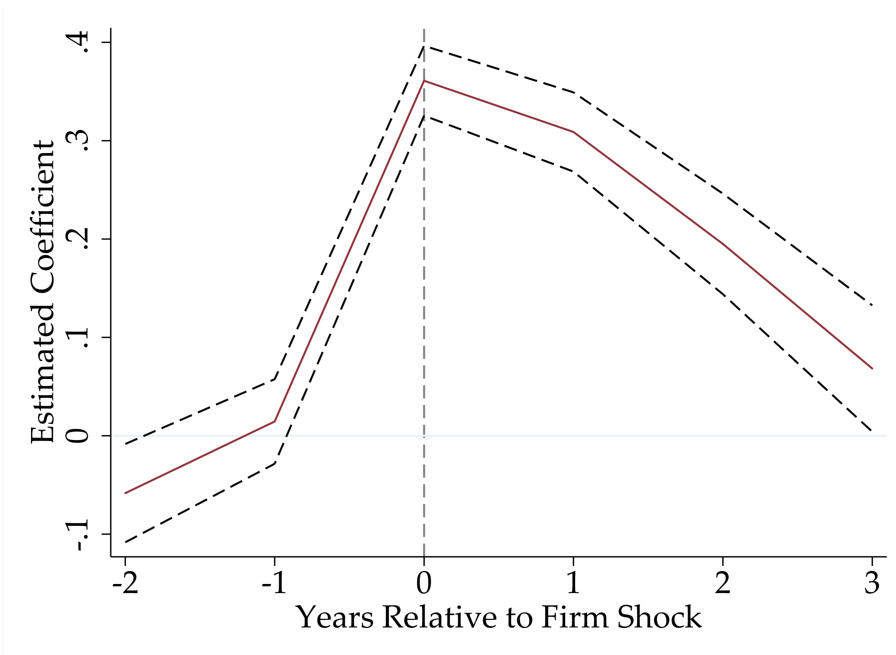
Table 2.14: **Bank Factors Extraction - Old Shocks, New GIV**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Dependent Variable: Bank Return on Loans (RoA)							
	OLS		Instrumented with GIV					
	Pooled	Pooled	Pooled	Positive	Negative	Pooled	Positive	Negative
(1) $GIV^{u^1}_{i,t}$	0.118	0.125	0.100	0.111	0.170	0.094	0.120	0.121
	(0.027)	(0.026)	(0.039)	(0.148)	(0.100)	(0.032)	(0.119)	(0.088)
(2) $GIV^{u^2}_{i,t}$	0.118	0.125	0.123	0.168	0.272	0.100	0.122	0.221
	(0.027)	(0.026)	(0.052)	(0.156)	(0.246)	(0.044)	(0.127)	(0.222)
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bank Controls	No	Yes	No	No	No	Yes	Yes	Yes

Notes: This table reports results from regressing bank-level return on loans on bank-level aggregated firm shocks $\bar{\tilde{\epsilon}}_{j,t}$, instrumented by two alternative Granular IVs. Rows (1)-(2) are based on firm shocks $u^1_{i,j,t}$ and $u^2_{i,j,t}$ obtained from running PCA separately on each bank's sample of borrowers using equation (2.26). These are the residuals remaining after extracting 1 and 2 common components, respectively, at the bank level. Standard errors (in parentheses) are clustered at the bank level.

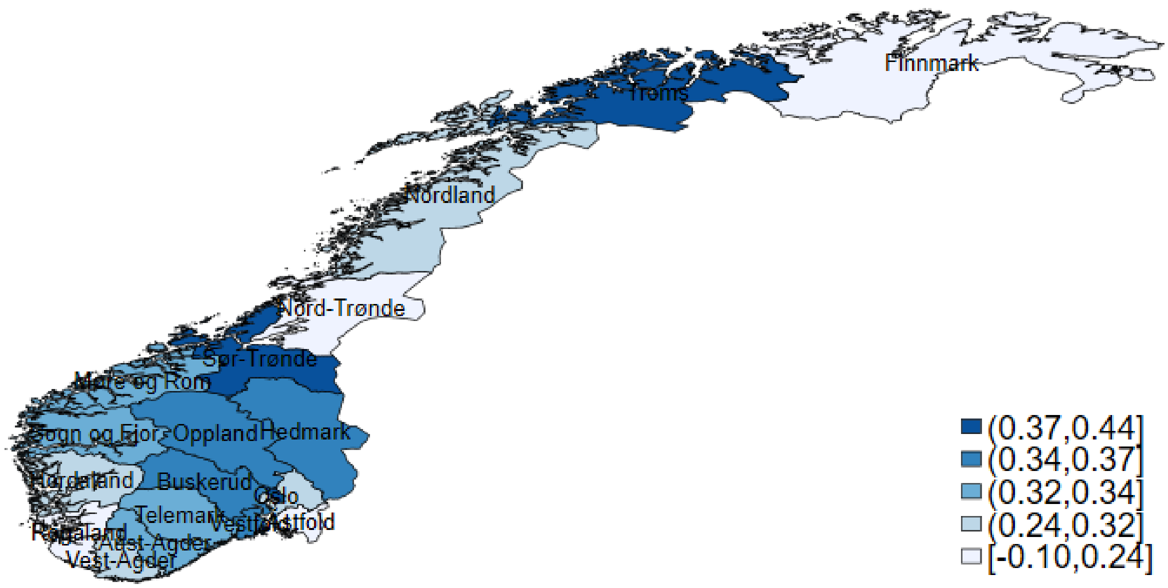
2.10 Additional Loan-Level Results

Figure 2.5: **Loan Outcomes by Horizon**



Notes: This graph plots results in the form of an event study where we regress loan-level returns on leads and lags of the idiosyncratic firm shock. Coefficients are plotted by horizon (in years) of the dependent variable. Specifications are based on equation (2.2). Firm shocks are estimated based on specification (2.1). Dashed lines are 95% confidence bands.

Figure 2.6: **Geographical Origins of Granular Credit Risk**



Notes: This picture is a colored map of 19 administrative counties (*fylke*) of Norway. Each shade of blue represents the county-specific strength of the pass-through from idiosyncratic firm shocks to return on loans. These correspond to county-specific slope shifters (slope dummies) introduced into the main loan regression 2.2. Shapefiles are from the Norwegian Mapping Authority (*Kartverket*).

Table 2.15: **Loan Outcomes - Firm Balance Sheet Heterogeneity**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline	0.361							
	(0.018)							
Shock x Low Leverage$_{t-1}$		0.345						
		(0.020)						
Shock x High Leverage$_{t-1}$		0.450						
		(0.047)						
Shock x High Assets$_{t-1}$			0.345					
			(0.018)					
Shock x Low Assets$_{t-1}$			0.976					
			(0.170)					
Shock x High Equity$_{t-1}$				0.352				
				(0.020)				
Shock x Low Equity$_{t-1}$				0.410				
				(0.044)				
Shock x Long Debt Duration$_{t-1}$					0.289			
					(0.020)			
Shock x Short Debt Duration$_{t-1}$					0.753			
					(0.046)			
Shock x Low Bank Reliance$_{t-1}$						0.314		
						(0.022)		
Shock x High Bank Reliance$_{t-1}$						0.497		
						(0.031)		
Shock x High Credit Rating$_{t-1}$							0.250	
							(0.025)	
Shock x Low Credit Rating$_{t-1}$							0.483	
							(0.026)	
Shock x Old Firms$_{t-1}$								0.313
								(0.020)
Shock x Young Firms$_{t-1}$								0.576
								(0.041)
Bank x Industry x Year x Loan-type x County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Additional controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	292825	292825	292825	292825	292825	292825	292825	292825
R^2	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167

Notes: This table reports results from loan-level regressions of loan returns on idiosyncratic firm shocks interacted with various lagged firm characteristics. Each characteristic is a dummy which takes the value of 1 for firms which are in the highest decile of leverage (defined as equity over assets), share of bank credit to total credit, and share of short-term debt to total debt; firms in the lowest deciles of total assets and total equity; firms with an below-A credit rating; and firms younger than 3 years. Baseline is the pooled estimation from Table 2.2. Standard errors (in parentheses) are double clustered at the firm-year level.

Table 2.16: **Loan Outcomes - Extensive Margin**

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Return on Loan				
	Baseline	Firm Exit	Firm Entry	Firm Bankruptcy	Ever Bankrupt
Firm Shock	0.361	0.387	0.322	0.365	0.360
	(0.018)	(0.019)	(0.019)	(0.018)	(0.019)
Exit / Entry / Bankruptcy		0.613	-1.707	0.699	0.572
		(0.075)	(0.073)	(0.161)	(0.079)
Interaction		-0.259	0.260	-0.133	0.014
		(0.067)	(0.059)	(0.133)	(0.068)
Full Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	292825	292825	292825	292825	292825
R^2	0.167	0.167	0.169	0.167	0.167

Notes: This table reports estimates from loan-level regressions of loan returns on firm shocks interacted with firm entry, exit, and bankruptcy dummies. Firm entry (exit) dummies equal 1 for firms which entered (exited) the year before (following) the firm shock. Firm bankruptcy is a dummy that equals 1 for firms which declare bankruptcy the year following the firm shock. Ever bankrupt is a dummy that equals 1 for firms which have *ever* declared bankruptcy during the 2003-2015 period, and not necessarily directly following the firm shock. All specifications include Bank x Firm Industry x Year x Loan-type x Firm County interacted fixed effects. Standard errors (in parentheses) are double clustered at the firm-year level.

Table 2.17: **Loan Outcomes - Firm Ownership Heterogeneity**

	(1)	(2)	(3)	(4)	(5)
	Dependent Variable: Loan RoA				
	All Firms	Private Firms	State Firms	Community Firms	Financial Vehicles
Firm Shock	0.335	0.336	0.478	0.089	1.145
	(0.016)	(0.019)	(0.654)	(0.120)	(0.966)
Bank x Year x County FE	Yes	Yes	Yes	Yes	Yes
Observations	330490	234074	162	2526	389
R^2	0.051	0.053	0.243	0.282	0.214

Notes: This table reports estimates from loan-level regressions of loan returns on firm shocks originating from firms with different ownership structure. Each column presents results from a specification in which only that particular ownership type is included. Numbers of observations do not add up because many firms are not assigned ownership classifications. Standard errors (in parentheses) are double clustered at the firm-year level.

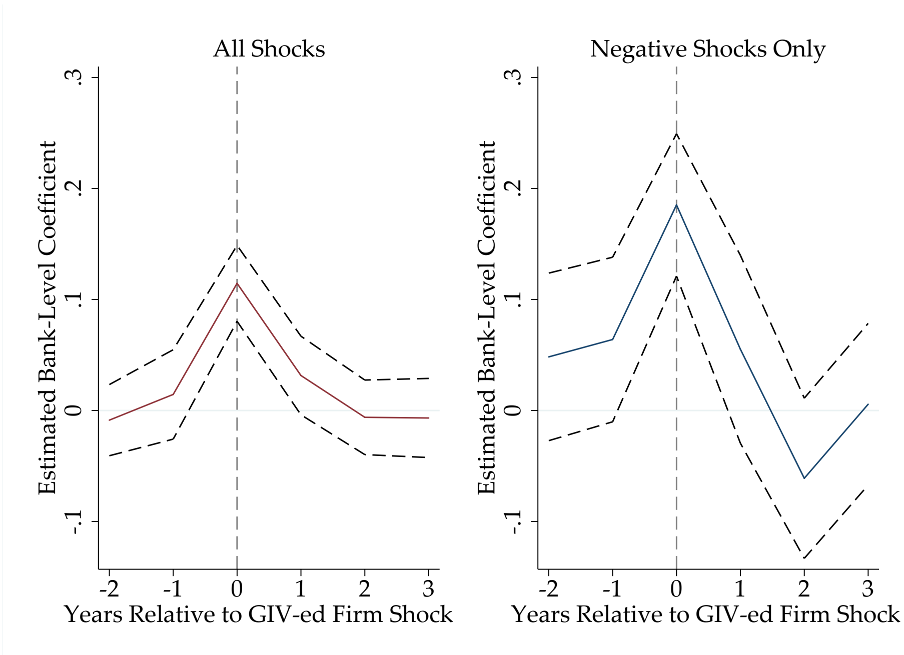
Table 2.18: **Loan Outcomes - Firm Industry Heterogeneity**

	(1)	(2)	(3)	(3)	(4)	(5)
	Dependent Variable: Return on Loan					
	All Firms	Manufacturing	Mining	Construction	Real Estate	Agriculture
Firm Shock	0.335	0.356	0.401	0.414	0.064	0.215
	(0.016)	(0.050)	(0.251)	(0.040)	(0.034)	(0.055)
Bank x Year x County FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	330490	34232	1097	60169	8531	7773
R^2	0.051	0.091	0.364	0.082	0.197	0.201

Notes: This table reports estimates from loan-level regressions of loan returns on firm shocks coming from firms from different sectors. Each column presents results from a specification in which firms from only that particular sector are included. Mining includes petroleum industries. Numbers of observations do not add up because many firms are not assigned industry classifications. Standard errors (in parentheses) are double clustered at the firm-year level.

2.11 Additional Bank-Level Results

Figure 2.7: **Bank Outcomes by Horizon**



Notes: This figure plots results in the form of an event study where we regress bank-level returns on leads and lags of the bank-level aggregated firm shock $\bar{\epsilon}_{i,t}$ instrumented by the granular IV. Specifications are based on equation (2.7). The GIV is constructed following equation (2.6). The left panel includes all shocks, and the right panel includes negative shocks only ($\bar{\epsilon}_{i,t} < 0$). Coefficients are plotted by horizon (in years) of the dependent variable. Dashed lines are 95% confidence bands.

Table 2.19: **Bank Loan Portfolio Writedowns and Sharpe Ratio**

	(1)	(2)	(3)	(4)
	Writedowns		Sharpe Ratio	
Granular Credit Shock	-0.016	-0.015	0.057	0.052
	(0.009)	(0.011)	(0.069)	(0.037)
Bank FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Bank Controls	Yes	Yes	Yes	Yes
Instrumented by GIV	No	Yes	No	Yes
Observations	1184	1184	1206	1206
R^2	0.937	0.071	0.654	0.025

Notes: This table reports results from regressing bank-level (log) loan writedowns and the Sharpe ratio on bank-level aggregated firm shocks $\bar{\epsilon}_{i,t}$. Columns (1) and (3) are OLS on equation (2.3), while in columns (2) and (4) the aggregated shocks are instrumented by the granular IV as in equation (2.7). The GIV is constructed following equation (2.6). Bank controls include lagged bank total assets, leverage, liquidity, number of loans, deposit to assets ratio and financial assets to assets ratio. Standard errors (in parentheses) are clustered at the bank level.

Table 2.20: **Bank Outcomes - Heterogeneity**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Bank Shock x Low Risk Weights	0.104						
	(0.042)						
Bank Shock x High Risk Weights	0.137						
	(0.040)						
Bank Shock x Low RWA		0.173					
		(0.037)					
Bank Shock x High RWA		0.029					
		(0.036)					
Bank Shock x Low Capital Ratio			0.090				
			(0.040)				
Bank Shock x High Capital Ratio			0.134				
			(0.039)				
Bank Shock x Low Loan HHI				0.068			
				(0.040)			
Bank Shock x High Loan HHI				0.138			
				(0.039)			
Bank Shock x Low Number of Loans					0.135		
					(0.046)		
Bank Shock x High Number of Loans					0.090		
					(0.030)		
Bank Shock x Low Liquidity						0.095	
						(0.045)	
Bank Shock x High Liquidity						0.135	
						(0.038)	
Bank Shock x Low Profitability							0.109
							(0.045)
Bank Shock x High Profitability							0.126
							(0.037)
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1208	1208	1208	1211	1211	1211	1211
R^2	0.101	0.106	0.101	0.103	0.102	0.102	0.101

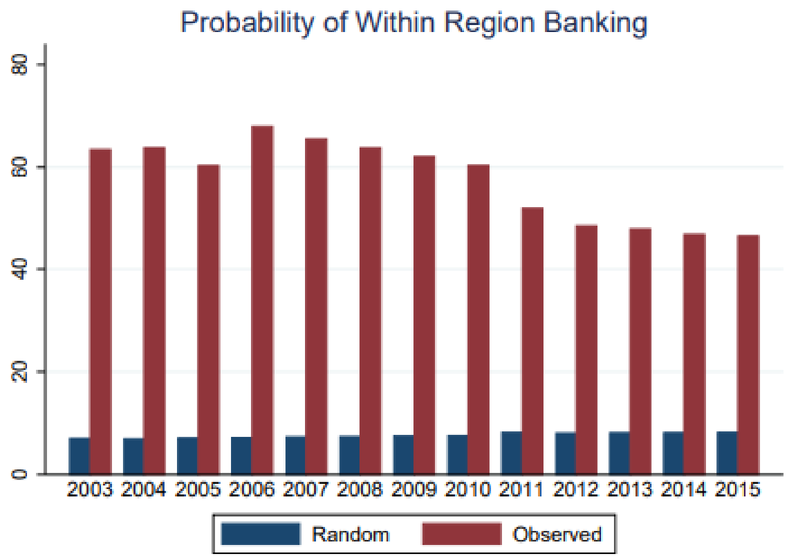
Notes: This table reports results from regressions of bank-level returns on corporate loans on GIV-instrumented idiosyncratic shocks interacted with lagged bank characteristics. In all columns, characteristics are cut based on the 50th percentile. Risk weights are obtained by dividing risk-weighted assets (RWA) by book assets. The regulatory capital ratio is defined as regulatory capital over RWA. Loan HHI refers to the within-bank Herfindahl index of loan concentration. Liquidity is defined as cash holdings over book assets. Profitability is defined as profit before taxes over book assets. All specifications include the following bank controls: lagged total assets, leverage, liquidity, number of loans, deposit to assets ratio, and financial assets to total assets ratio. Standard errors (in parentheses) are clustered at the bank level.

Table 2.21: **Bank Outcomes - Inspecting the Risk-Taking Channel**

	(1)		(2)		(3)	
	Low RWA	High RWA	Low CapRatio	High CapRatio	Low HHI	High HHI
Bank Shock x Low Risk Weights	0.156	-0.005	0.070	0.119	0.056	0.117
	(0.050)	(0.058)	(0.054)	(0.051)	(0.067)	(0.048)
Bank Shock x High Risk Weights	0.212	0.061	0.108	0.168	0.075	0.187
	(0.070)	(0.039)	(0.058)	(0.062)	(0.051)	(0.063)
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1208	1208	1208	1208	1208	1208
R2	0.105	0.105	0.101	0.101	0.103	0.103
	(4)		(5)		(6)	
	Low NumLoans	High NumLoans	Low Liquid	High Liquid	Low Profit	High Profit
Bank Shock x Low Risk Weights	0.120	0.079	0.060	0.131	0.114	0.086
	(0.063)	(0.043)	(0.076)	(0.048)	(0.057)	(0.057)
Bank Shock x High Risk Weights	0.162	0.105	0.127	0.149	0.095	0.163
	(0.065)	(0.045)	(0.047)	(0.074)	(0.049)	(0.059)
Bank FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1208	1208	1208	1208	1208	1208
R2	0.102	0.102	0.101	0.101	0.101	0.101

Notes: This table reports results from regressions of bank-level returns on corporate loans on GIV-instrumented idiosyncratic shocks, double interacted with bank risk weights and additional characteristics. In all specifications, characteristics are cut based on the lagged 50th percentile. For example, column (1) presents estimates for banks with low risk weights and low risk-weighted assets (RWA), low risk weights and high RWA, high risk weights and low RWA, and high risk weights and high RWA. Similarly for all other columns. Risk weights are obtained by dividing risk-weighted assets (RWA) by book assets. The regulatory capital ratio (CapRatio) is defined as regulatory capital over RWA. HHI refers to the within-bank Herfindahl index of loan concentration. NumLoans refers to the (log) number of loans in the portfolio. Liquid refers to the liquidity ratio, defined as cash holdings over book assets. Profit refers to the profitability ratio, defined as profit before taxes over book assets. All specifications include the following bank controls: lagged total assets, leverage, liquidity, number of loans, deposit to assets ratio, and financial assets to total assets ratio. Standard errors (in parentheses) are clustered at the bank level.

Figure 2.8: **Home Bias in Within-Region Banking**



Notes: This figure shows the extent to which there is home-bias in the Norwegian corporate credit market. Source: Juelsrud and Wold (2020). Specifically, red bars show the *observed* fraction of loans within a given year in our sample where the firm and the bank are located in the same county (within-region loans). The blue bars show the counterfactual share of within-region loans, where we assume random matching between firms and banks. Given random matching, the probability that a firm i borrows from a bank j operating in that region is the sum of the aggregate/national market share of bank j.

2.12 Narratives

In this section we validate our baseline idiosyncratic firm shock $\epsilon_{j,t}$ with a narrative-based approach. It is important to confirm that $\epsilon_{j,t}$ truly reflect economically meaningful information about firm performance. We focus on the bottom 1st percentile of realizations of $\epsilon_{j,t}$ in the final shock distribution used in our analysis and search through the Norwegian news media for corresponding narratives.[40] In a lot of cases, some of which are outlined below, we find that our idiosyncratic shock matches actual, sizable economic events.

One of the most adverse shocks in our sample was experienced by Hera Vekst - a waste management company - in 2008. For that year, we estimate an unexpected idiosyncratic shock $\epsilon_{j,t}$ of -1.39, corresponding to approximately an unexpected drop in value added of -139%. This drop was seemingly generated by the sudden closure of the company's main facility, enforced by local authorities. Local authorities enforced the closure due to the company's repeated violation of air pollution standards. According to local news reports, the smell from the waste management facility was "far in excess of what the local population should tolerate" (nrk.no, 2011).

The company Nergard Sild, a mid-sized herring farmer, experienced an idiosyncratic shock $\epsilon_{j,t}$ of -1.2 in 2010 according to our estimates. National news reports attributed this loss to over-investment in a processing facility for herring (nrk.no, 2012). The investment had been planned in 2009 "when the quota was 1 million tons." Once the realized quota turned out to be much smaller than expected (370,000 tons), Nergard Sild closed down the processing facility, leading to substantial losses.

Staying in the domain of fish farming, another major shock in our sample is for the company Wilsgard Fiskeoppdrett. Wilsgard Fiskeoppdrett - a fish farming company specializing in salmon - experienced an idiosyncratic shock of -1.23 in 2015. According to national media, the reason for this drop was a massive outbreak of salmon lice (iLaks.no, 2015). The outbreak was so severe that the Norwegian Food Safety Authority threatened the company with a daily fine until the situation got under control, worrying that the outbreak would spread along the coast.

Subaru Norge AS - the lead importer of Subaru in Norway - had an idiosyncratic shock of -1.21 in 2007 according to our estimates. The drop was generated by a tax hike on gasoline-fueled cars, which changed the relative price on gasoline-fueled vs. diesel-fueled cars. While the tax was levied on all gasoline-fueled cars, Subaru was the only major brand without a viable diesel alternative (DN, 2007). As a consequence, the number of new cars sold for Subaru dropped from 3800 to 344 cars by August the following year.

The horticulture company F.Dalene Gartneri AS had an idiosyncratic shock of -1.17 in 2008. According to local news media, the manager of the company was engaged in substantial fraud,

[40] The 1st percentile of the idiosyncratic shock distribution is -.905, while the 5th percentile is -.459.

which culminated in arson on the main facility to recoup an insurance premium of approximately 5 million USD (pd.no, 2011).

Fraud is the reason for another one of the most negative shocks in our sample. FIBO - an aluminum producer - experienced an idiosyncratic shock of -1.25 in 2007 according to our estimates, which ultimately lead to their subsequent bankruptcy in 2009. The bankruptcy trustee had substantial criticism towards the board of the company, going far in pointing to outright fraud and stating that the case was so severe that its "report would and should be sent to the Financial Supervisory Authority for further study" (jarlsbergavis.no, 2011).

Next, consider the case study of the furniture producer Ekornes, which in 2015 had an estimated idiosyncratic shock of -1.24. The company blamed adverse conditions in the German consumer market, one of their largest client bases. Looking for the causes, the CEO of Ekornes pinpointed the uncertain economic environment and the conflict between Russia and Western Europe. "Germans are careful. They save in bad times. The conflict between Western Europe and Russia has affected Germans more than in Norway" (e24.no, 2014).

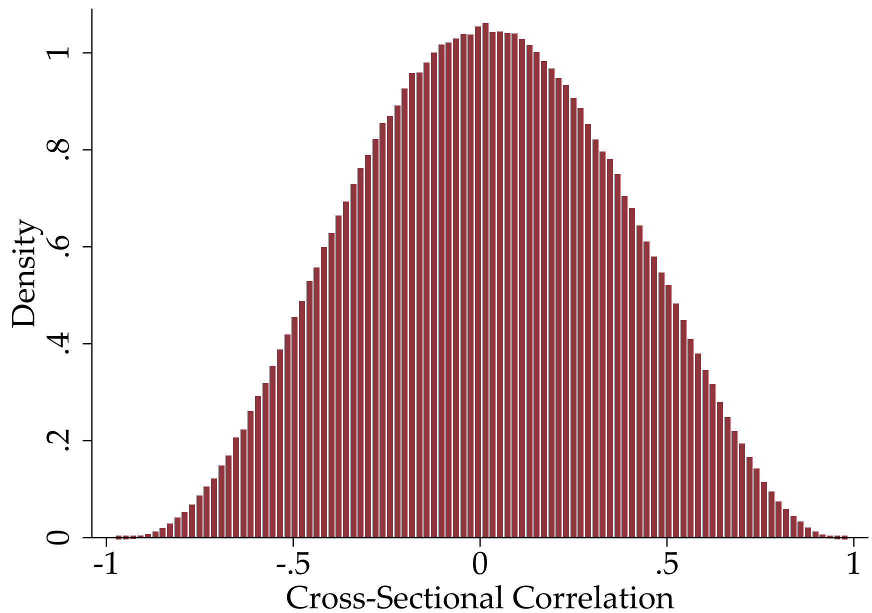
Other notable shocks in our sample include the shipping company Volstad Shipping, which in 2008 experienced an idiosyncratic shock of -1.28 due to misplaced foreign currency positions (smp.no, 2012), and the company Bergen Group Intech which in 2010 experienced an idiosyncratic shock of -1.33 due to under-performance of their investments in Iceland. Those assets were subsequently sold due to "not being within the core areas of the company" (Finansavisen, 2011).

Our estimated shocks also pick up less dramatic events. For instance, consider the firm GC Rieber Oils, a firm specializing in producing Omega 3-based products. In 2013, they recorded an $\epsilon_{j,t}$ of -0.24. The incident which caused this, according to local newspapers, was an accidental spill of between 500 and 800 litres of raw material from the company's factories into the local harbor (Naeringsliv, 2013). The spill was eventually managed and dealt with thanks to the local municipality and fire services. The spill lead to "substantial economic losses" for the company, according to the CEO (Naeringsliv, 2013).

2.13 Robustness Tests and Auxiliary Findings

Figure 2.9: **Pairwise Cross-Sectional Correlation of Firm Shocks**

Panel A: Histogram



Panel B: Summary Statistics

	Number of Pairs	Mean	Std. Dev.	Min	Max
Firm Shock	1,861,485	0.019	0.342	-0.977	0.985

Notes: These figures report all pairwise cross-sectional correlation coefficients for idiosyncratic firm shocks. The sample includes a balanced panel of firms over 2003-2015. Panel A presents the histogram, and Panel B reports summary statistics. Firm shocks are extracted based on specification 2.1.

Table 2.22: **Robustness to the Great Financial Crisis**

	(1)	(2)	(3)	(4)	(5)	(6)
	Loan-Level			Bank-Level		
Firm Shock	0.361	0.432	0.322	0.117	0.091	0.108
	(0.018)	(0.032)	(0.022)	(0.029)	(0.051)	(0.037)
All Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Bank Controls				Yes	Yes	Yes
Observations	292825	102879	189946	1211	472	737
R^2	0.167	0.158	0.172	0.101	0.066	0.127

Notes: This table reports timing robustness for baseline loan- and bank-level regressions from Tables 2.2 and 2.3, respectively. Columns (1)-(3) report results of loan and columns (4)-(6) of bank outcomes, respectively. Columns (1) and (4) are baseline estimates. Columns (2) and (5) include only the pre-2009 period. Columns (3) and (6) include only the post-2009 period.

Table 2.23: **Placebo Regressions - Permutation Tests**

	Simulations	True Coefficient	Event Frequency	Event P-value
		Loan Outcomes		
Permuted Firm Shock	1000	0.361	0	0.000
		Bank Outcomes		
Permuted Firm Shook, Pooled	1000	0.116	0	0.000
Permuted Firm Shock, Positive Only	1000	0.016	838	0.838
Permuted Firm Shock, Negative Only	1000	0.194	0	0.000

Notes: This table reports results from Monte Carlo permutation regressions where loan or bank return on loans are regressed on firm shocks that are randomly permuted. The last two rows report results when permuted shocks are positive or negative only, respectively. Columns report the number of simulations, the true coefficients based on Table 2.2 column (3) and Table 2.3 columns (3)-(5), the number of events where permutations produced estimates that are as large as the true estimate (in absolute value) by chance, and the associated p-values.

Table 2.24: **Placebo Regressions - Random Shocks**

	Number of Draws	Mean	Std. Dev.	Min	Max
		Loan Outcomes			
Placebo Firm Shock	1000	0.001	0.007	-0.018	0.021
		Bank Outcomes			
Placebo Firm Shook, Pooled	1000	0.000	0.005	-0.016	0.018
Placebo Firm Shock, Positive Only	1000	0.001	0.018	-0.053	0.049
Placebo Firm Shock, Negative Only	1000	-0.000	0.014	-0.041	0.046

Notes: This table reports results from a placebo exercise where loan or bank outcomes are regressed on sequences of randomly generated numbers. In each row, placebo shocks are randomly drawn from the interval of the true shock. The last two rows report results when shocks are positive or negative only, respectively. Columns report the number of random draws and summary statistics of the regression coefficients: mean, standard deviation, minimum, and maximum.

Table 2.25: **Estimating Fixed Effect Linear Models with AR(1) Disturbances**

	Borrower Level	Bank Level	Firm Industry Level	County Level
Autoregressive Coef.	0.318	0.122	0.241	0.223
Standard Deviation	0.267	0.107	0.254	0.251

Notes: This table reports parameter estimates of a linear unbalanced panel fixed effects model with a disturbance that follows an autoregressive process of order 1. Estimates for the autoregressive coefficient and the standard deviation of the error term are reported. Columns report results for various levels of aggregation. Idiosyncratic firm shocks are extracted based on specification 2.1 and then aggregated to different levels with loan shares as weights.

Table 2.26: **Theoretical Model Parameter Estimates**

Firm Size	Parameters			Loan Distribution Variance
	α	λ	$\alpha\tau$	
Sales	1.26	1.005	1.388	Infinite
	(0.002)	(0.548)	(0.413)	
Assets	1.321	0.923	1.587	Infinite
	(0.001)	(0.361)	(0.887)	
Equity	1.495	1.086	1.641	Infinite
	(0.002)	(0.467)	(1.144)	

Notes: This table reports estimates of key parameters of the model described in Section 2.7. α, λ and $\alpha\tau$ represent the Pareto power parameter of the firm size distribution, the firm's debt demand elasticity, and the sufficient statistic of the Singh-Maddala distribution, respectively. Standard errors (standard deviations for λ and $\alpha\tau$) are in parentheses.

References

BAI, J. AND S. NG (2002): "Determining the number of factors in approximate factor models," *Econometrica*, 70, 191–221.

DN (2007): "Subaru salget stuper," *https://www.dn.no/subaru-salget-stuper/1-1-982883*.

E24.NO (2014): "Ekornes sier de har sovet i timen," *https://e24.no/boers-og-finans/ekornes-sier-de-har-sovet-i-timen/23274867*.

FINANSAVISEN (2011): "Bergen Group kvitter seg med to selskaper," *https://finansavisen.no/nyheter/boers-finans/2011/07/bergen-group-kvitter-seg-med-to-selskaper*.

GABAIX, X. AND R. KOIJEN (2020): "In search of the origins of financial fluctuations: The inelastic markets hypothesis," *Manuscript*.

ILAKS.NO (2015): "Trues med dagmulkt," *https://ilaks.no/trues-med-dagmulkt/*.

JARLSBERGAVIS.NO (2011): "Kritikk etter Fibo-konkursen," *https://www.jarlsbergavis.no/nyheter/kritikk-etter-fibo-konkursen/s/1-67-5793472*.

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

NAERINGSLIV, D. (2013): "Bandak-konsernet overende," *https://www.tk.no/nyheter/slapp-ut-over-500-liter-fiskeolje/s/1-113-6707592*.

NRK.NO (2011): "Slutt for Hera Vekst," *https://www.nrk.no/innlandet/slutt-for-hera-vekst-1.7792540*.

——— (2012): "Nergård legger ned i Bø," *https://www.nrk.no/nordland/nergard-legger-ned-i-bo-1.8834498*.

PD.NO (2011): "Gartnerens store tragedie," *https://www.pd.no/lokale-nyheter/gartnerens-store-tragedie/s/1-89-5759714*.

SMP.NO (2012): "Kjempeår for Volstad Shipping," *https://www.smp.no/nyheter/article240019.ece*.

STOCK, J. H. AND M. W. WATSON (2016): "Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics," in *Handbook of macroeconomics*, Elsevier, vol. 2, 415–525.

Chapter 3

Bewley Banks

This chapter develops a non-linear, quantitative macroeconomic model with heterogeneous monopolistic financial intermediaries, incomplete markets, default risk, endogenous bank entry, and aggregate uncertainty. The model generates a bank net worth distribution fluctuation problem analogous to the canonical Bewley-Huggett-Aiyagari-Imrohoglu environment. Our framework nests Gertler and Kiyotaki (2010) (GK) and the standard Real Business Cycle model as special cases. We present four general results. First, relative to the GK benchmark, banks' balance sheet-driven recessions can be significantly amplified, depending on the interaction of endogenous credit margins, the cyclicality of a precautionary lending motive and the (counter-) cyclicality of intermediaries' idiosyncratic risk. Second, equilibrium responses to aggregate exogenous shocks depend explicitly on the conditional distributions of bank net worth and leverage, which are endogenous time-varying objects. Aggregate shocks to banks' balance sheets that hit a concentrated and fragile banking distribution cause significantly larger recessions. A persistent consolidation in the U.S. banking sector that matches the one observed over 1980-2020 generates a large economic contraction and an increase in financial instability. Third, we document, and match, novel stylized facts on both the cross-section of credit margins and the cyclical properties of the first three moments of the cross-sectional distributions of financial intermediary assets, net worth, leverage, loan margins, and default risk. We find that shocks to capital quality and to leverage constraint tightness ("financial shocks") can match fluctuations in the U.S. financial sector very well. Finally, we use the model to identify and characterize episodes of systemic banking crises. Such events are associated with large economic recessions, spikes in bank leverage, and large drops in the number of intermediaries.

We thank Marco Bellifemine for excellent research assistance. Jamilov thanks the AQR Asset Management Institute and the Wheeler Institute for Business and Development for financial support.

3.1 Introduction

The 2007-2008 Global Financial Crisis has transformed the way the profession thinks about the role of financial intermediaries in the economy. A large, new literature that followed has recognized the importance of financial frictions in a rich variety of setups that span constraints on risk-taking, imperfect competition, credit cycles, and moral hazard. However, the vast majority of existing papers are studies of the *first moment* rather than of the full *distribution* of financial intermediaries. In this paper, we lay out a tractable, quantitative macroeconomic framework with a banking sector where the time-varying distributions of financial intermediary net worth and leverage are at the core of the analysis. We supplement our quantitative analysis with novel stylized facts on both the cross-section and the cyclicality of different moments of the U.S. banking distribution.

Our model builds on the work of Jamilov (2020) and introduces aggregate uncertainty and heterogeneous banks into an environment with monopolistic competition in financial intermediation, incomplete markets, default risk, and endogenous bank entry. Bank credit is intermediated through a time-varying mass of local credit markets. Each market features a unique financial variety or amenity that is desired by the household. A single bank intermediates all markets and can charge market-specific rate margins. The elasticity of substitution across credit markets is constant trough time and states of nature, in a way similar to the goods market structure in Dixit and Stiglitz (1977) or Melitz (2003). Imperfect competition is the source of an aggregate credit supply externality, in the spirit of Blanchard and Kiyotaki (1987): the bank does not internalize the impact of market-specific loan margins on aggregate investment demand.

In addition to monopolistic financial intermediation, we assume that the bank faces local, partially uninsurable idiosyncratic rate of return risk in the spirit of Benhabib et al. (2018). Idiosyncratic risk, jointly with credit market power, creates a banks' net worth distribution fluctuation problem analogous to the canonical Bewley-Huggett-Aiyagari-Imrohoglu environment (Bewley, 1977; Huggett, 1990; Aiyagari, 1994; Imrohoglu, 1996). Importantly, our modelling approach eliminates *scale invariance*: all dynamic choices in the financial sector depend on bank-specific characteristics such as the level of net worth.[1] The number of local credit markets is determined in equilibrium through endogenous entry, similarly to the heterogenous non-financial firms model of Melitz (2003). Equilibrium yields a non-trivial, dynamic distribution of bank assets. The presence of aggregate risk makes this distribution, in principle an infinitely-dimensional object, a relevant "state variable". Aggregate state-dependency on the distribution is thus achieved explicitly, a result that is not feasible in other environments that feature scale invariance and complete markets. Under

[1] Eliminating scale invariance is a crucial step that separates our paper from the rest of the literature where a "representative" intermediary is the commonplace assumption. An important exception is Coimbra and Rey (2019) who study ex-ante heterogeneous banks. In contrast, our model delivers ex-post heterogeneity in returns and bank size due to market incompleteness and loan market power.

perfectly competitive credit markets, and in the absence of idiosyncratic risk, our "Bewley Banks" setup nests the canonical Real Business Cycle model and the Gertler and Kiyotaki (2010); Gertler and Karadi (2011); Gertler et al. (2016) macro-banking frameworks as special cases ("GK" models henceforth).[2] Such high tractability allows us to conduct a variety of benchmarking exercises.

A key advantage of our Bewley Banks framework is that we can target cyclical properties of *higher-order* moments of any banking characteristic. To that end, we document a comprehensive set of stylized facts on the distribution of U.S. financial intermediaries, both in the cross-section and over the business cycle. In particular, we focus on the mean, dispersion, and concentration of intermediaries' assets, net worth, leverage ratio, loan margins, and default risk. Nailing down empirical moments of the bank distribution turns out to be a non-trivial task. There is multi-modality in the aggregate data and rich heterogeneity across *sub-industries* of the broader financial sector. We therefore provide additional sets of industry-specific facts on depository institutions, brokers and dealers, insurance companies, etc. We also report new stylized facts for other developed economies including Australia, Canada, France, Germany, and the United Kingdom.

As in the data, our model generates a right-skewed ergodic distribution of banks' asset and leverage, and an inverse relationship between credit margins and bank size. The model also generates business cycle statistics that approximate the cyclical properties of the different moments of the U.S. banking distribution rather well. Relative to the GK benchmark, and in response to banks' balance sheet ("capital quality") shocks, our baseline model generates equilibrium dynamics for key aggregate variables that are considerably dampened. This result is due to two reinforcing channels. First, a precautionary lending motive, due to market incompleteness, makes each intermediary accumulate more equity capital than in the GK counterfactual. Greater assets and net worth, coupled with lower aggregate leverage in the precautionary stochastic steady state, leave the financial sector in a less fragile initial condition when aggregate shocks hit. Second, credit market power is an additional margin of adjustment in response to adverse shocks that allows the bank to boost profits in high marginal-utility states by raising prices and reducing quantities by less. As a result of both mechanisms, aggregate contractions get dampened.

In order to counteract the precautionary lending motive, we allow idiosyncratic risk to be state-dependent and *counter-cyclical*. We provide first-pass non-parametric evidence in favor of this channel using the full distribution of U.S. bank returns over the past 20 years. In recessions, the dispersion (skewness) of both transitory and persistent bank return shocks rises (falls). This counter-cyclicality of bank returns is particularly striking for the ongoing COVID-19 pandemic.[3] Motivated by this empirical finding, we then suppose that in low aggregate states a larger mass

[2]To enable the cleanest possible comparison, we solve GK with fully non-linear methods as well.

[3]Among many others, Bloom et al. (2018) document that non-financial microeconomic risk rises in recessions. Guvenen et al. (2014) provide similar evidence for the case of household income risk.

of credit markets can experience low idiosyncratic return draws. In this case, the direct impact of business cycle fluctuations on bank balance sheets dominates the ex-ante precautionary lending motive. Recessions get amplified considerably, both in terms of real macroeconomic aggregates and in the financial sector. Notably, the number of active intermediaries falls by an order of magnitude more than in the acyclical risk counterfactual. Overall, in the Bewley Banks framework crises could be either amplified or dampened depending on whether idiosyncratic bank risk is counter-cyclical or not.

Macroeconomic and financial dynamics with Bewley Banks are driven by the interaction of *three main forces*: credit market power, idiosyncratic return risk, and endogenous entry. The tractability of the model allows us to isolate the differential contribution of each force sequentially. From this exercise we gain three main insights. First, shutting down credit market power considerably amplifies the aggregate sensitivity to exogenous shocks. Second, without idiosyncratic return risk the economy is less risky and less responsive to aggregate shocks. Third, eliminating endogenous entry dampens the responsiveness by a negligible amount. In the baseline model with acyclical idiosyncratic risk, the large dampening role of credit market power dominates the amplifying effect of market incompleteness. The extensive margin is muted and does little to affect the responses of either output or consumption.

An important result of the paper pertains to the general aggregate *state dependency* on the dynamic distributions of bank net worth and leverage. In the model, equilibrium responses of aggregate output and consumption to exogenous aggregate shocks depend explicitly on the conditions in the banking cross section. A negative bank capital quality shock that hits the economy in a state with a more concentrated and fragile distribution of net worth and leverage generates a significantly larger cumulative loss in output, consumption, bank assets, and considerably greater levels of bank leverage and loan margins.

Recent papers by Jamilov (2020) and Corbae and D'Erasmo (2020) document a multifold increase in both the degree of dispersion and concentration in the U.S. banking sector over the 1980-2020 period, coupled with a steady decline in the absolute number of depository institutions. We explore the role of second moment shocks as a source of business cycle fluctuations. We find that positive transitory shocks to the dispersion of bank assets, the magnitude of which matches the data, have large negative effects both on the real economy and the financial sector. Specifically, aggregate output, consumption, bank assets and net worth each fall by substantial amounts whereas bank leverage, credit margins, and default risk all rise considerably. We implement this particular exercise by allowing the agents in the model to explicitly track and forecast higher-order moments of the distribution of bank net worth. This computational approach follows the original idea in the seminal works by Krusell and Smith (1996, 1998). That is, we keep track of the first two moments of the banking distribution as an approximation for what is otherwise an infinitely-dimensional object.

In general, we find that persistent shocks to higher-order moments of the banking distribution have a significant and lasting effect. The key for this result is that our model generates concentrated, right-skewed distributions of bank assets and leverage.

Another avenue that we explore in the paper is the identification and characterization of systemic *banking crises* using event study methods. We apply tools from the open-economy macroeconomics literature that looks at financial crises in emerging economies (Mendoza, 2010). We simulate a long time-series from our model and define an economic crisis as an episode with unusually low measured TFP. We then collect all instances of such events and compute period-specific averages of key macro-financial aggregates. We find that economic crises occur in conjunction with banking crises. In relative terms, banking crises in the Bewley economy lead to more dampened aggregate contractions than in the GK counterfactual with perfect competition and no idiosyncratic risk. Crises in the Bewley economy are also less financially unstable - i.e., bank leverage increases by less. This is a variant of the financial competition-stability trade-off (Hellman et al., 2000). Imperfect credit-market competition acts as a buffer against shocks to financial stability but at the price of high loan margins and a greater steady-state loss in consumer welfare.

Crises in the Bewley Banks economy with *counter-cyclical* idiosyncratic risk, however, are significantly amplified. These episodes correspond to greater contractions in output, consumption, bank assets and net worth. Furthermore, the number of active intermediaries falls by an order of magnitude more than in the baseline model with acyclical idiosyncratic risk. In terms of financial stability, the financial sector is more fragile than in the baseline: aggregate leverage and default risk both increase by a greater amount. Interestingly, crises in the Bewley Banks economy with counter-cyclical risk are characterized by a persistent *decline* of credit margins in the build-up periods. Meanwhile, crises in the baseline case with acyclical idiosyncratic risk are associated with a *rise* of credit margins in the the same build-up phase. This is an interesting testable implication of our model. Generally speaking, this result suggests that competition and market power in the financial sector could be leveraged for the ex-post measurement of financial conditions as well as forward-looking diagnostics and forecasts of distress and crises.

Up to this point, the sole source of aggregate uncertainty in all quantitative experiments has been a so-called capital quality shock. In the final exercise, we take business cycle moments in the data as given and run a horse race across six different types of aggregate shocks with the purpose of matching as many unconditional correlations as possible. We consider aggregate shocks to Hicks-neutral total factor productivity, quality of aggregate capital, intermediary dividend payouts, credit markups, leverage constraint tightness, and degree of market incompleteness. For each shock type, we solve our model under the assumption that it is the only source of aggregate uncertainty in the environment. We find that shocks to capital quality and to leverage constraint tightness ("financial shocks") can match fluctuations in the U.S. financial sector very well.

Solving our model numerically is potentially a challenging task for at least four reasons. First, without scale invariance the distribution of financial intermediary assets, an infinitely-dimensional object, is now a "state variable" that needs to be kept track of. We solve this issue with a variant of the Krusell and Smith (1998) algorithm, where we assume that the bank forms linear forecasts for a limited set of moments of the cross-sectional distribution. Along these lines, the bank must also rationally forecast future aggregate prices (loan rates) which are set at the level of a local credit market. The two forecasts taken together allow the bank to pin down the projected return on aggregate capital. Second, the market for deposit holdings must clear at each point of the aggregate state space. Third, dynamics of the cross-sectional bank distribution must be consistent with the dynamic problem of the incumbents *and* new entrants. Finally, the intermediary faces an occasionally binding constraint on leverage. Importantly, this constraint may bind on any part of the idiosyncratic or aggregate state space. In Section 3.3.13 we describe our computational algorithm in detail.

Related Literature. This paper builds on the recent work by Jamilov (2020) and Corbae and D'Erasmo (2019, 2020) who explore aggregate implications of the rise of U.S banking concentration. De Loecker et al. (2020) and Diez et al. (2018) document a large, 50%+ increase in markups (margins) in the U.S. financial industry over 1980-2016. The apparent consolidation of market power in the financial sector calls into question the assumption of perfect competition that is made in most existing quantitative macro-banking models. Gerali et al. (2010) and Cuciniello and Signoretti (2015), among others, studied the role of imperfect competition in general equilibrium macroeconomic models. Our paper is different in our assumption of persistent ex-post heterogeneity, which allows us to achieve a smooth equilibrium ergodic distribution of bank leverage. Our financial intermediaries are also exposed to insolvency-driven default risk, which is priced into a distribution of deposit rates.

Among studies that explore banking industry dynamics in general equilibrium, Corbae and D'Erasmo (2019) is the paper that is closest to ours but uses a different approach.[4] Authors focus on dynamic capital requirements in a quantitative model of oligopolistic financial competition with dynamic interactions between one large bank with market power and many small perfectly-competitive institutions. We differ from Corbae and D'Erasmo (2019) in at least three main respects. First, our approach to modelling imperfect banking competition follows a large literature that works with CES aggregation (Dixit and Stiglitz, 1977; Melitz, 2003). This approach is highly tractable and "portable" - our monopolistically competitive banking bloc can be readily enhanced to

[4]Banking industry dynamics have been explored in several other recent papers including Capelle (2019), Rios Rull et al. (2018), Nguyen (2015), Christiano and Ikeda (2013), Davydiuk (2020), Martinez-Miera and Repullo (2010). This broad literature builds on the first generation of mostly partial-equilibrium models on the financial competition-stability trade-off (Allen and Gale, 1998; Hellman et al., 2000; Allen and Gale, 2004).

allow for nominal rigidities or open-economy features[5]. Second, in our model local credit market-specific actions are not internalized in the aggregate. In the oligopolistic banking setup of Corbae and D'Erasmo (2019), actions of the "lead bank" and "fringe banks" are fully internalized. Our modelling approach yields a powerful aggregate credit supply externality, which acts as an additional channel of amplification on firm investment demand. Jamilov (2020) shows that internalization of the credit supply externality moves steady-state aggregates such as welfare, output, and bank credit by double-digit percentage points. Finally, our model nests explicitly the canonical Real Business Cycle model and the GK models as special cases, which adds to its tractability. Reverting to GK entails simply a re-calibration of three structural parameters.

We also build on the important work by Coimbra and Rey (2019) who study the impact of ex-ante financial heterogeneity on systemic macroeconomic risk and financial stability. Their model features both intensive and extensive margins, similarly to ours. Our approach differs from theirs in several substantive ways. First, we achieve persistent "ex-post" financial heterogeneity due to incomplete markets and exposure to idiosyncratic risk, while Coimbra and Rey (2019) assume ex-ante heterogeneity in Value-at-Risk constraints. Second, in our setup, trading markets are incomplete and the banking distribution is a time-varying, endogenous state variable which must be kept track of. Our model also, importantly, allows for credit market power and a mismatch between the cost of funds and the price of bank credit. Finally, the number of active intermediaries in our model is a time-varying object that moves with the business cycle.[6]

Our assumption of idiosyncratic rate of return risk in banking is not ad-hoc or far-stretched. Several recent studies document that intermediaries are not perfectly insured against non-systematic shocks at various layers of aggregation. Galaasen et al. (2020) show that financial intermediaries in Norway are exposed to idiosyncratic borrower-level risk at multiple layers of aggregation. In particular, they show that bank portfolios are highly regionally concentrated due to regional "home bias" in lending and that local market-specific risk is hard to hedge. Their finding maps closely to our modelling approach where idiosyncratic risk is tied to the spatial distribution of locally differentiated credit markets. Paravisini et al. (2020) find that persistent specialization of banks by export market leaves them vulnerable to idiosyncratic shocks originating with foreign partner-countries. Agarwal et al. (2020) find that banks which over-exposed to the Mexican energy sector were much more likely to suffer from the industry-specific negative shock of 2014. Although rigorously micro-founding idiosyncratic risk is beyond the scope of our quantitative framework, our parsimonious modelling approach can capture and operationalise the general idea reasonably well.

[5]We extend our framework to include nominal rigidities and a role for monetary policy in Jamilov and Monacelli (2020).

[6]Other papers such as Korinek and Nowak (2016), Boissay et al. (2016), Goldstein et al. (2020), Begenau and Landvoigt (2018) also work with equilibrium models of ex-ante heterogeneity in the financial sector.

The rest of the paper is structured as follows. Section 3.2 reports stylized facts on the banking distribution in the cross-section and over the business cycle. In Section 3.3, we lay out our model. Section 3.4 describes our calibration strategy, shows the model policy functions and ergodic distributions, and demonstrates the responsiveness to aggregate fluctuations. Section 3.5 inspects the model mechanism by isolating each key moving part. Section 3.6 presents our main quantitative results and experiments. Section 3.7 explores different types of exogenous aggregate shocks. Finally, Section 3.8 concludes.

3.2 Stylized Facts on the Distribution of Financial Intermediaries

In this section we document key stylized facts on the banking industry in the cross-section and over the business cycle. In Section 3.2.2 we report facts on the cross-sectional distribution. In Section 3.2.3 we look at the behavior of the banking distribution over the business cycle. In Section 3.2.4 we study cyclicality of idiosyncratic bank return shocks. We focus on the whole financial sector in main text and discuss sub-industry heterogeneity in Appendix 3.9.2. We also report, subject to data availability, relevant statistics for non-U.S. countries in Appendix 3.9.3. Our data Appendix 3.9.1 provides further details on raw data and our data cleaning approaches. Alternative robustness checks are documented in Appendix 3.9.4.

3.2.1 Data Description

Variable Definition

For every financial characteristic that we describe below, we are interested in computing the first three moments of the time-varying cross section. We begin with the first moment which we proxy with the (unweighted) mean μ_t of a generic panel x_{jt} with N_t being the size of the population:

$$\mu_t(x) = \sum_{j}^{N_t} \frac{1}{N_t} x_{jt}$$

For the second moment, we compute the time-varying standard deviation σ_t:

$$\sigma_t(x) = \sqrt{\frac{\Sigma_j^{N_t}\left(x_{jt} - \mu_t(x)\right)^2}{N_t}}$$

For the third moment, depending on the characteristic, we compute either the Herfindahl index or the statistical skewness. We use the HHI primarily for bank assets and net worth, and skewness for the leverage ratio. The Herfindahl H_t of x_{jt} is defined using the usual formula, where s_{jt} is the share of bank j in market x in time t:

$$H_t(x) = \sum_{j}^{N_t} s_{jt}(x)^2$$

Finally, for statistical skewness S_t we use the Pearson's standardized third moment:

$$S_t(x) = \frac{\frac{1}{N_t}\sum_j^{N_t}\left(x_{jt} - \mu_t(x)\right)^3}{\left[\frac{1}{N_t-1}\sum_j^{N_t}\left(x_{jt} - \mu_t(x)\right)^2\right]^{\frac{3}{2}}}$$

Bank balance sheets

Our main source of bank balance sheet information is the Compustat database. We start by extracting financial intermediary assets and net worth for the financial sector in the U.S. Our definition of (book) leverage is the ratio of bank assets to net worth. The aggregate distribution of intermediary leverage is multi-modal, because different sub-industries of the broader sector have heterogeneous business models. We therefore also report statistics for six sub-sectors: depository credit institutions, non-depository credit institutions, brokers and dealers, insurance companies, real estate companies and brokers, and holding companies and investors. We define these sectors based on the SIC classification. We also work with Compustat Global, which has information for non-U.S. institutions. Existing data for non-U.S. developed economies is, however, limited. Our main aggregate time-series for assets, net worth, and leverage runs from 1985q1 until 2020q1. We explore samples with alternative starting dates in Appendix 3.9.4.

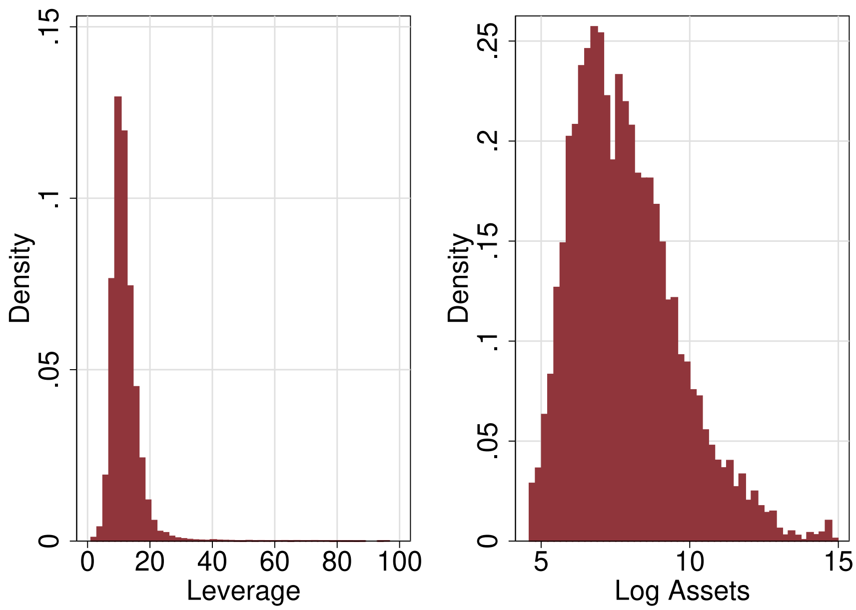
Loan margins

We construct our own, bottom-up measure of credit margins. To that end, we employ the Compustat Banks dataset. Our baseline definition of a credit margin is the bank-level ratio of Total Interest and Related Income over Total Interest and Related Expenses. Our approach is similar to Corbae and D'Erasmo (2019), who measure credit margins by the ratio of returns over the marginal cost of funds. Our main loan margin series runs from 1993q1 until 2020q1, and refers to depository institutions. For each quarter, as with the balance sheet data, we compute the mean, standard deviation, and skewness based on the raw bank-level panel data.

Bank default risk

We proxy bank-level default risk with data on Credit Default Swaps (CDS) provided by Markit. Our baseline measure is the 5-year CDS spread, which is the most liquid. Raw data is available at daily frequency, and we aggregate it to the bank-quarter level. We were only able to compute

Figure 3.1: **Banking Distributions in the Data**



Notes: Histograms of bank leverage and (log) total assets in the US data. Leverage is defined as book assets over book equity. Source: Compustat.

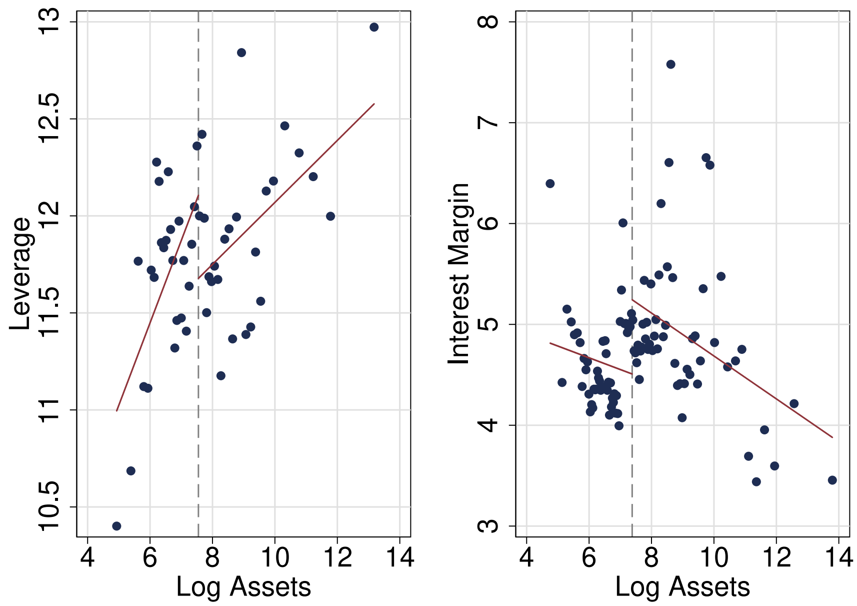
CDS spreads for the financial sector as a whole; industry-level information is not available. Our measures are available for both U.S. and non-U.S. based institutions. The final dataset runs over 2002q1-2020q1. For each quarter, we calculate the mean, standard deviation, and skewness of the distribution of CDS spreads.

Each variable, unless explicitly noted otherwise, is logged and linearly detrended. Our main proxy of the "business cycle" is real U.S. Gross Domestic Product (GDP), which is detrended and seasonally adjusted. The same filtering is performed on the model-simulated time-series as well, so the model and the data are directly comparable.

3.2.2 Banking in the Cross Section

We begin by first documenting facts on the banking distribution by analyzing the cross-section. In Figure 3.1 we plot the unconditional histograms of bank leverage (defined as the ratio of book assets over book equity) and the log of assets. We observe that both distributions are highly right-skewed. The vast majority of intermediaries have a leverage ratio somewhere in the 5-15 region. However, there is a small mass of highly risky banks. An important feature of our structural model will be the ability to generate such a skewed distribution thanks to the combination of a persistent idiosyncratic return process and market incompleteness.

Figure 3.2: **Bank Size, Leverage and Interest Margins**



Notes: Binned scatterplots of (log) assets against bank leverage and interest margins. Leverage is defined as book assets over book equity. Interest margins are defined as the ratio of total interest income to total interest and related expenses. Vertical dashed lines represent discontinuity at the median of log assets. Red lines of fit are separate lines of linear fit for values above and below the discontinuity. See main text for further details. Source: Compustat.

In Figure 3.2 we plot binned scatterplots of bank size against book leverage and our measure of the interest margin. We define bank size as total book assets, which is also in line with a similar empirical exercise done in Coimbra and Rey (2019). We construct these plots in three steps. First, we residualize both y-axis and x-axis variables from the time fixed effect in order to absorb any common time-varying factors. Second, we construct 100 equally-sized bins of log assets. Each bin has at least 300 observations. For each bin we compute unweighted averages of log assets, book leverage, and the interest margin. Finally, we fit the data points separately for "small" and "large" intermediaries, which we separate based on the median of the distribution of assets. This discontinuity on the graph is captured by a dashed vertical line.

From Figure 3.2 we document two empirical regularities. First, the conditional correlation of book leverage and bank size is positive. Larger banks are also more levered, in line with the findings in Coimbra and Rey (2019). Second, the conditional correlation of bank size and interest rate margins is negative. It is statistically significant at the 5% level. Larger intermediaries have lower interest rate margins, on average. The relationship is particularly strong for banks in the top half of the size distribution.

3.2.3 Banking over the Business Cycle

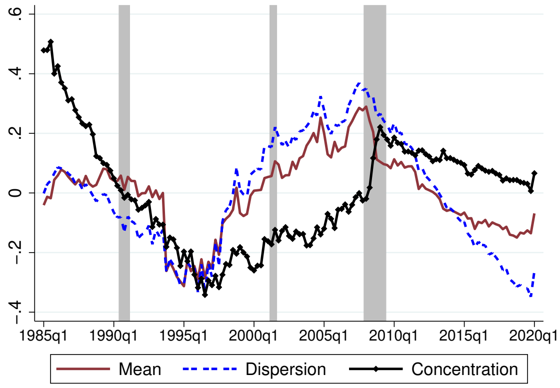
Next we present a series of facts documenting the cyclical properties of the first three moments of the banking distribution.

Figure 3.3 plots the time-series of the first three moments of financial intermediary assets, net worth (equity), leverage, margins, and default risk (CDS spreads). The underlying sample is for the U.S. only. In panel (a) we report the mean, standard deviation, and HHI of total intermediary *assets*. We see clearly from the picture that average assets are pro-cyclical, which is consistent with observations in Nuno and Thomas (2016). The second moment is also (highly) pro-cyclical. The third moment, however, is counter-cyclical. In panel (b) we plot the time-series for bank *equity*. Cyclical properties of the first three moments of bank equity are the same: positive, positive, and negative.

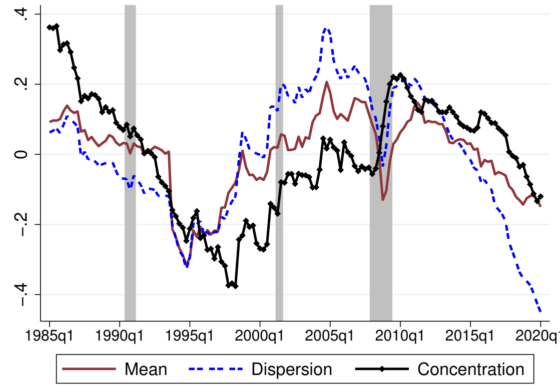
In panel (c) we report statistics on the bank *leverage ratio* (defined as book assets over book equity). The first moment is highly pro-cyclical, consistent with the evidence in Adrian and Shin (2010, 2011). It is important to note, however, that there is substantial variation in risk-taking management across different types of financial intermediaries, a point we will return to again in Appendix 3.9.2. We find that the second moment of leverage is slightly pro-cyclical and the third moment (skewness) is strongly counter-cyclical. Overall, for the total financial sector the mean and dispersion of bank assets, net worth, and leverage are pro-cyclical while concentration is highly counter-cyclical. Counter-cyclicality of the third moment appears to be a robust feature of the data.

In panel (d) we plot the first three-moments of the distribution of *loan margins*. We find that average margins are strongly counter-cyclical, and so is the dispersion of margins. Concentration, however, is very pro-cyclical. Consistently with our results, Corbae and D'Erasmo (2019) also document that average credit margins are counter-cyclical but they do not explore higher-order moments like us. Finally, in panel (e) we document properties of *CDS spreads*. The first two moments of the distribution are counter-cyclical, a fact that is largely consistent with canonical theory. The third moment is slightly pro-cyclical but this result is sensitive to sample selection.

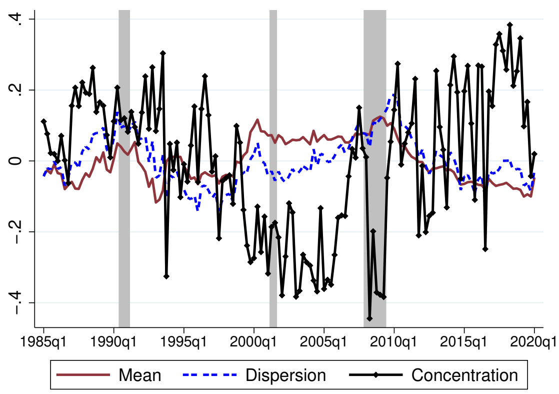
Figure 3.3: **Cyclical Components of the Distribution of U.S. Financial Intermediaries**



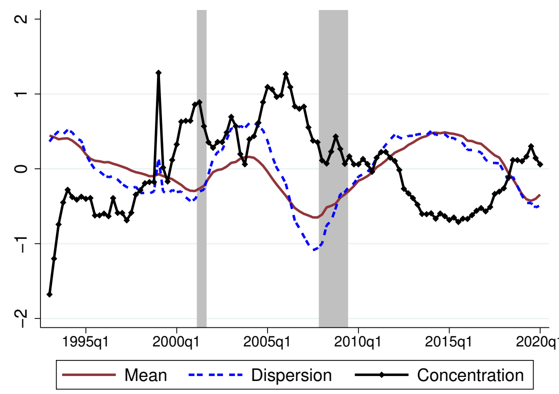
(a) Assets



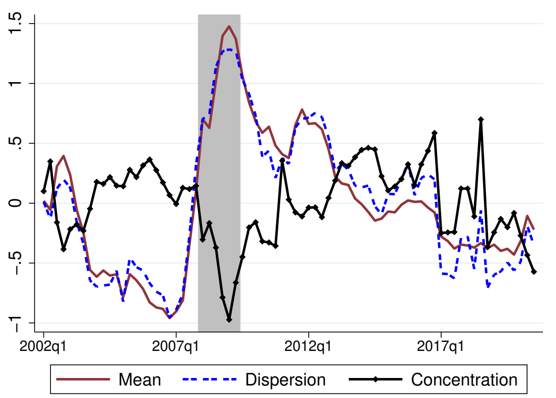
(b) Equity



(c) Leverage



(d) Credit Margins



(e) CDS Spreads

Notes: Every variable has been logged (except the skewness of leverage) and linearly detrended. Shaded areas represent US recessions based on the NBER classification. Bank balance sheet data is from Compustat. CDS data is from Markit. See Appendix 3.9.1 for variable definitions and further details.

Table 3.1 displays a summary of our banking business cycle facts. Hence we see that intermedi-

Table 3.1: **Business Cycle Statistics - Aggregate U.S. Data**

	Mean of	Dispersion of	Concentration of
Correlation			
Assets - GDP	0.498	0.642	-0.568
Net Worth - GDP	0.211	0.544	-0.472
Leverage - GDP	0.701	0.043	-0.641
Margins - GDP	-0.563	-0.370	0.725
CDS Spreads - GDP	-0.325	-0.309	0.033
Standard Deviation (%)			
Assets	13.383	19.371	18.281
Net Worth	11.268	18.076	16.640
Leverage	6.036	6.855	20.157
Margins	31.046	42.404	56.595
CDS Spreads	57.751	58.496	32.021

Notes: Dispersion is measured as time-varying standard deviation. Concentration is measured with skewness. For every variable except CDS spreads the sample is 1985q1:2020q1. For CDS spreads the sample is 2002q1:2020q1. Every variable has been logged (except the skewness of leverage) and linearly detrended. Bank balance sheet data is from Compustat. CDS data is from Markit. See Appendix 3.9.1 for variable definitions and further details.

ary balance sheet quantities have a pro-cyclical mean and dispersion, and a strongly counter-cyclical HHI. Book leverage is pro-cyclical in the first two moments, and counter-cyclical in the third moment. The mean and standard deviation of credit margins and CDS spreads are both counter-cyclical, while skewness is pro-cyclical. In the lower panel of Table 3.1 we report standard deviations of time-series fluctuations of our main variables. The results can be summarized into two key facts. First, the volatility of balance sheet quantities - assets, net worth, leverage - is far smaller than the volatility of either credit margins or default risk. Second, higher-order moments appear to be more volatile than the mean. In particular, measures of concentration are especially volatile.

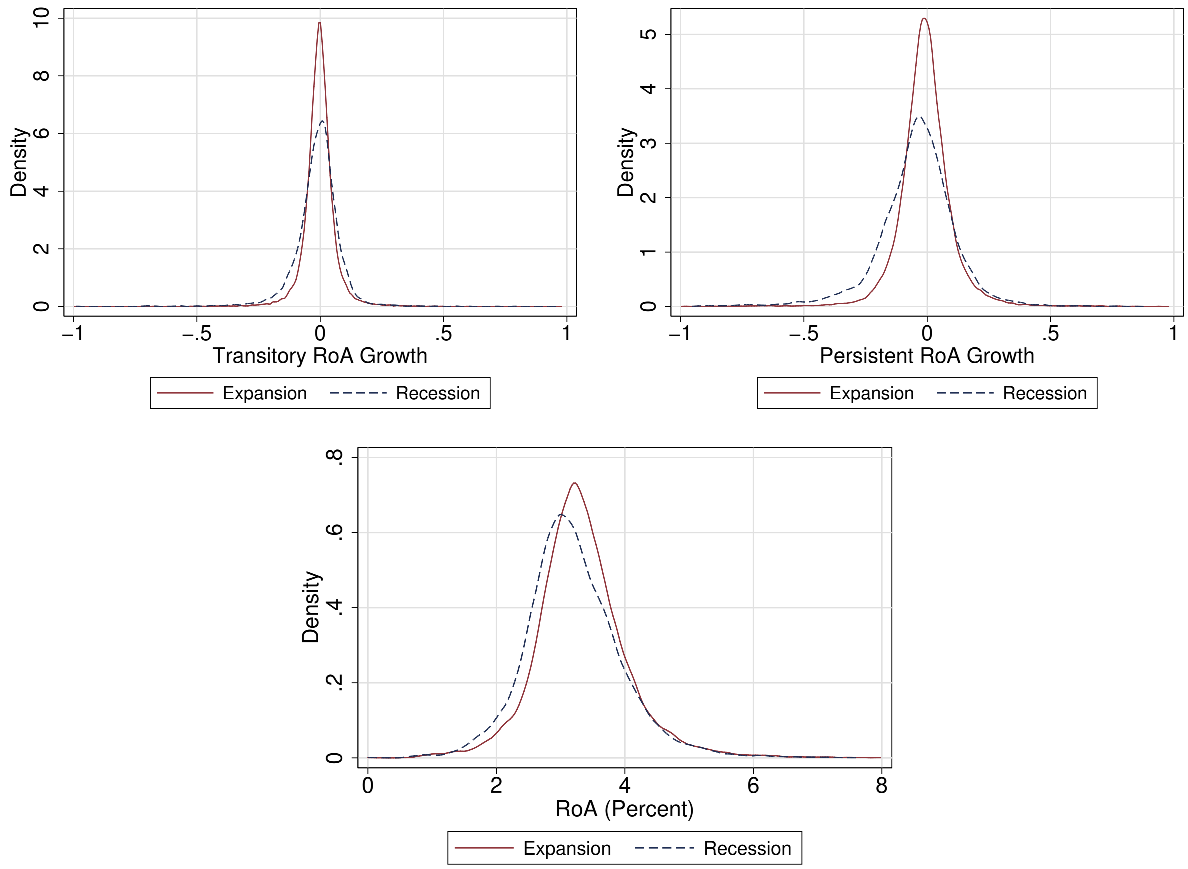
3.2.4 Cyclicality of Bank Returns

We conclude this section by analyzing micro-banking uncertainty over the business cycle. To measure uncertainty, we compare the full distribution of bank returns in expansions and recessions. We define returns at the level of a bank as returns on assets (RoA). RoA is our favorite measure because it takes into account the amount of debt or risk involved in creating those returns. Our RoA index is defined as the ratio of net income to total assets.[7]

We focus not only on the RoA in levels but also in terms of growth over short and long horizons.

[7] Replacing net income with total interest income doesn't change the results as the two series are strongly correlated.

Figure 3.4: **Cyclicality of Bank Returns**



Notes: Densities of bank returns on assets (RoA), defined as net income scaled by total assets. Transitory and persistent return shocks are defined as $y_t - y_{t-1}$ and $y_t - y_{t-4}$, respectively, where y_t is bank-level (log) RoA. Recessions follow NBER definitions and include the 2001-2002 crisis, the Great Recession, and the COVID-19 Pandemic which corresponds to the first two quarters of 2020. Expansions are quarters which are not included into the recession sample. The sample is 2000:q1-2020:q2 and includes only US commercial banks. Source: Compustat.

To that end, we look at one- and four-quarter RoA growth rates (specifically, $\log(\text{RoA})_t - \log(\text{RoA})_{t-j}$ for $j = 1, 4$). The two definitions roughly correspond to transitory and persistent shocks to bank returns and follow the non-parametric approach of Guvenen et al. (2014). Our sample construction follows closely Section 3.2.3 with the exception of us focusing now on the 2000:q1-2020:q2 period. Our sample captures three recessions, as defined by the National Bureau of Economic Research (NBER) classification: the 2001-2002 crisis, the Great Recession, and the COVID-19 pandemic which corresponds to the first two quarters of 2020.

The first set of results is depicted in Figure 3.4 and summarized in Table 3.2. Three observations are worthy of note. First, the mean of each of the three measures of bank returns falls in recessions. Second, in recessions the standard deviation rises for both transitory and persistent shocks. Third, the skewness becomes more negative in recessions, i.e., the distribution is more left-skewed during

Table 3.2: **Bank Return Risk - Summary Statistics**

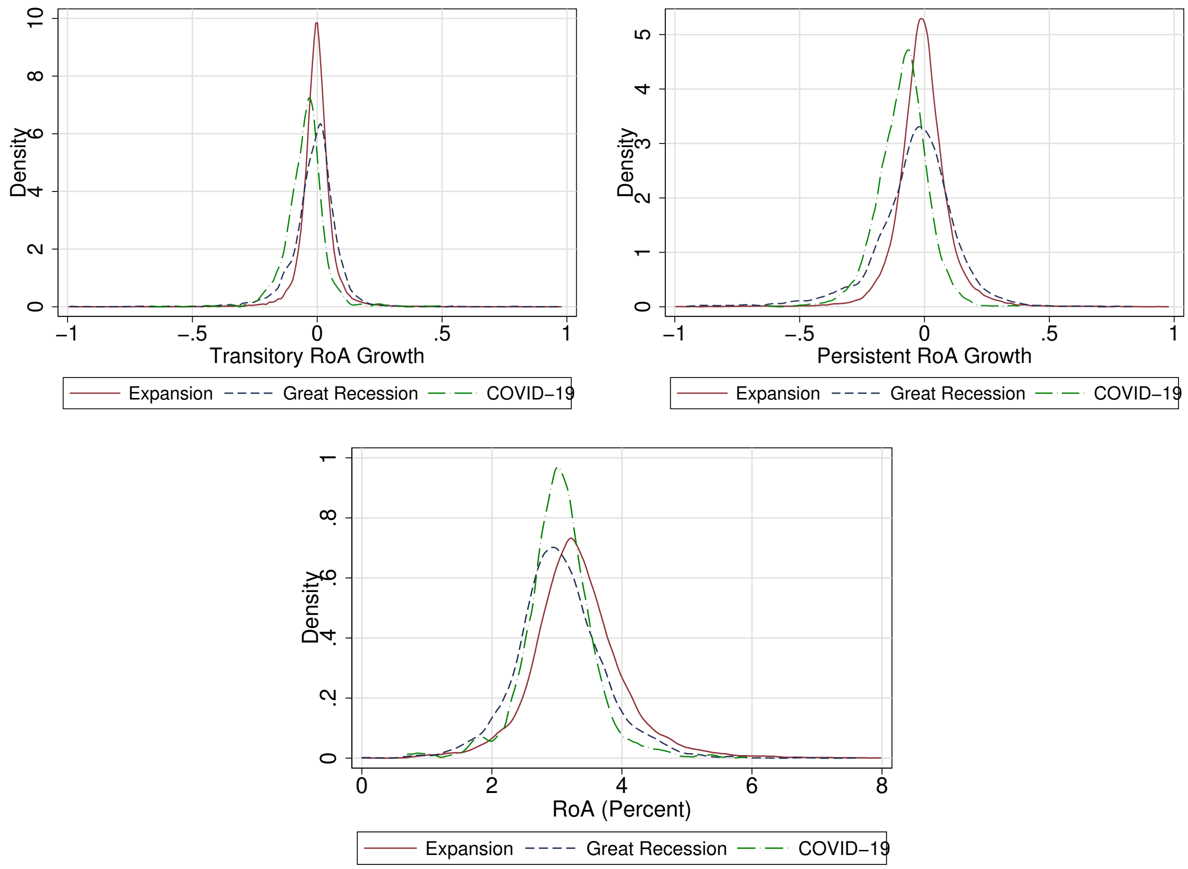
	Expansion			Recession		
	Transitory	Persistent	Level	Transitory	Persistent	Level
Mean	-0.0035	-0.0079	3.4062	-0.0108	-0.0425	3.2396
St Deviation	0.0738	0.1149	1.14	0.0994	0.1595	0.9502
Skewness	-0.4272	-0.1232	5.5641	-0.8572	-0.6295	4.1143

Notes: Summary statistics for the densities from Figure 3.4.

bad times. The latter two observations - counter-cyclicality of dispersion and pro-cyclicality of skewness - suggest that micro-banking uncertainty increases during recessions. This result is related to the established findings that microeconomic uncertainty (as approximated by various measures of income risk) is counter-cyclical for the case of both households (Guvenen et al., 2014) and firms (Bloom et al., 2018). Our contribution in this section is to provide first-pass non-parametric evidence for a similar channel, yet applied to financial intermediaries.

Counter-cyclical bank return risk can be further, and rather strikingly, illustrated for the ongoing COVID-19 pandemic. In Figure 3.5 we plot cross-sectional distributions of bank return shocks for the first two quarters of 2020. Relative to the distributions in expansions and even during the Great Recession, the left tail is currently very much stretched out as banks are more likely to experience large negative shocks to portfolio returns. As can be seen from the Figure, the pandemic is likely to have highly persistent effects as the dispersion of persistent returns shocks is especially large from a historical perspective. The considerable growth (decline) in dispersion (skewness) is driven by a severe underlying deterioration of firms' balance sheets and valuations which, in turn, are passed through to banks' balance sheets via bankruptcies or non-performing assets.

Figure 3.5: **Bank Returns during the Great Recession and COVID-19 Pandemic**



Notes: The COVID-19 Pandemic corresponds to the first two quarters of 2020. The sample is 2000:q1-2020:q2 and includes only US commercial banks. Source: Compustat.

3.3 Model

In this section we lay out our quantitative model with a dynamic, endogenous distribution of bank characteristics.

3.3.1 Overview

Time is discrete and infinite. The economy is populated by *four agents*: a representative household, a capital goods producer, a final goods producer, and a financial intermediary. The economy is subject to *aggregate uncertainty* in the form of shocks to the quality of aggregate capital. This shock proxies exogenous perturbations to the balance sheet of financial intermediaries. The role of the bank is to intermediate funds between a unit mass of identical households and productive capital. The bank is owned by the household and ultimately redistributes back all its accumulated wealth (net worth) via dividends. Dividends are paid out only upon exit. Every period, the bank finances its operations via accumulated net worth or through a nationwide retail market, where it obtains deposits from households. There are no wholesale funding markets available.

Bank credit is intermediated on a time-varying mass J_t of local credit markets. This mass is determined endogenously through entry and exit. A single financial intermediary intermediates funds across all of these markets. Credit markets are differentiated by unique local features such as amenities or variety in financial services. The elasticity of substitution between local credit markets, $\theta > 1$, is constant across time and aggregate states of nature. Imperfect substitutability across credit markets allows the intermediary to charge localized marked-up *margins* over the cost of funds.

The portfolio return of the banker consists of a systematic component and a persistent, idiosyncratic, *uninsurable* returns process. Idiosyncratic risk is market-specific and cannot be hedged because trading markets are incomplete. A long enough sequence of negative idiosyncratic return shocks can drive the bank into insolvency. There is no deposit insurance. Bank default risk gets competitively priced into the interest rates on deposits by the homogeneous, risk-averse household.

Entry into credit markets is endogenous. There is a large number of potential entrants - financiers - who become financial varieties conditional on entry. Entering financiers pay a fixed startup cost and obtain a one-time idiosyncratic return draw. Having observed the draw, financiers can either decide to operate or to immediately exit. The mass of entering financiers grows until bank profits (in expectation) remain above the startup costs. Involuntary exit occurs at rate $0 < \sigma < 1$.

Importantly, our model breaks down scale invariance, thereby generating a dynamic, endogenous cross-section of intermediary assets. The distribution of bank assets (loan portfolio), an infinitely dimensional object, is a new state variable in the model. We describe how we deal with the curse of dimensionality in Section 3.3.13. The equilibrium is also associated with the dynamic endogenous distributions of bank net worth, leverage, default risk, margins, and deposit rates.

3.3.2 Aggregate Technology

There is a continuum of perfectly competitive firms that produce the final good using an identical constant returns to scale Cobb-Douglas production function with capital and labor as inputs. Labor is supplied inelastically, for tractability. Output Y_t is the following function of aggregate capital K_t and labor L_t:

$$Y_t = AK_t^{\alpha} L_t^{1-\alpha} \tag{3.1}$$

with $0 < \alpha < 1$.

In the baseline model, the only source of aggregate uncertainty is a shock to the quality of capital ψ_t (Merton, 1973; Gertler and Kiyotaki, 2010; Gertler and Karadi, 2011). This shock captures fluctuations in the value of capital - its sudden obsolescence or valuation. We will explore alternative forms of aggregate uncertainty in the latter sections. We assume that ψ_t follows an AR(1) process.

$$\psi_{t+1} = (1-\rho_\psi) + \rho_\psi \psi_t + \epsilon^\psi_{t+1} \tag{3.2}$$

Capital accumulates over time according to the law of motion:

$$K_{t+1} = \psi_{t+1}\Big(I_t + (1-\delta)K_t\Big)$$

where I_t is aggregate investment of non-financial firms and $0 < \delta < 1$ is the constant depreciation rate. Wages are competitive and follow directly from the production function and firms' optimization. Return on aggregate capital R^k_t is:

$$R^k_{t+1} = \psi_{t+1}\left(\frac{A\alpha K^{\alpha-1}_{t+1} + (1-\delta)P_{t+1}}{P_t}\right) \tag{3.3}$$

which comprises profits and capital gains. The latter depend on the dynamics of the *aggregate* price of capital, P_t, which is determined in equilibrium by financial intermediary activities.

3.3.3 Local Credit Markets

There exists a time-varying mass J_t of local credit markets. The elasticity of substitution across markets is $\theta > 1$. Credit markets are differentiated by unique features that (local) borrowers derive utility from. Differentiated capital goods are assembled by a representative capital producing firm with a Dixit-Stiglitz aggregation technology from the mass J_t of available financial varieties $k_t(j)$ where $j \in [0, J_t]$.

$$K_t = \left[\int_0^{J_t} k_t(j)^{\frac{\theta-1}{\theta}} dj\right]^{\frac{\theta}{\theta-1}} \tag{3.4}$$

Financial variety-specific demand functions are obtained from the following maximization problem:

$$\max_{k_t(j)} \Big[K_t - \int_0^{J_t} p_t(j)k_t(j)dj\Big] \tag{3.5}$$

subject to technology (3.4). This yields the demand function for banks' lending activities:

$$k_t(j) = \left(\frac{p_t(j)}{P_t}\right)^{-\theta} K_t \tag{3.6}$$

where $p_t(j)$ is the price of capital in the local credit market (j) and P_t is the aggregate price of capital consistent with the competitive capital producing firm earning zero profits:

$$P_t := \left[\int_0^{J_t} p_t(j)^{1-\theta} dj\right]^{\frac{1}{1-\theta}} \tag{3.7}$$

3.3.4 Financial Intermediary

The monopolistic credit demand system (3.4)-(3.7) is taken as given by the intermediary. The intermediary starts each period with an initial stock of net worth $n \in \mathbf{N} \subset \mathbf{R}_+$ and must choose the stock of assets k(j), deposits d(j), and price of varieties p(j) while satisfying the balance sheet constraint:

$$d_t(j) + n_t(j) = p_t(j)k_t(j)^{\beta} \tag{3.8}$$

where $\beta > 1$ is a parameter that governs local decreasing returns to scale. The bank can borrow deposits d(j) from the household, subject to the market-specific interest rate $\bar{R}_t(j)$. The incumbent banker earns a portfolio return $R^T(j)$ that consists of the return on aggregate capital R^k - common across local markets - and an idiosyncratic component $\xi(j)$ which is specific to each local credit market:

$$R_t^T(j) = \kappa\xi_t(j) + (1-\kappa)R_t^k \tag{3.9}$$

Where $0 < \kappa < 1$ is a parameter that governs the ability to hedge local market-specific return risk. For tractability, assume that $\xi \in \Xi$ follows an AR(1) process:

$$\xi_t(j) = (1-\rho_\xi)\mu_\xi + \rho_\xi\xi_{t-1}(j) + \sigma_\xi\epsilon_t(j) \tag{3.10}$$

Let the finite state Markov representation of (3.10) be $\mathbf{G}_{\xi_{t+1},\xi_t}$. The law of motion of bank net worth is therefore:

$$n_{t+1}(j) = R_{t+1}^T(j)p_t(j)k_t(j) - \bar{R}_t(j)d_t(j) \tag{3.11}$$

Following Gertler and Karadi (2011) and Gertler and Kiyotaki (2010), the bank-household relationship is subject to a moral hazard friction. The bank has an exogenous incentive to divert bank assets. It has the capacity to divert no more than a fraction λ of the total value of local assets p(j)k(j). If considering to divert, the banker always manages to escape but the bank enters bankruptcy the following period. In equilibrium, it must be that the bank stays indifferent between operating honestly and diverting. This yields the following incentive constraint that puts a limit on bank leverage:

$$\lambda\ p_t(j)k_t(j) \leq\ V_t(j) \tag{3.12}$$

where $V_t(j)$ is the franchise value of the intermediary, whose recursion is defined below.

The market-specific probability of default is $\upsilon(j)$. Default risk is due to fundamental insolvency, i.e., when bank net worth at normal market prices is non-positive:

$$\upsilon_t(j) = \Pr\Big(n_{t+1}(j)(n_t(j), \xi_t(j)) \le 0\Big) \tag{3.13}$$

Conditional on insolvency, the household recovers only an endogenous fraction of promised payments $x_t(j)$, to be defined later. This risk is priced by the household through equilibrium deposit rates. Remaining assets get transferred to the capital producing firm who produces K_t as normal.

The distribution of financial varieties ("banks", for short) is summarized by the probability measure μ defined on the Borel algebra B that is generated by open subsets of the product space $\mathbf{B} = \mathbf{N} \times \Xi$, corresponding to the distribution of incumbent banks with net worth n and idiosyncratic return realization ξ. The aggregate state of the economy is (ψ, μ, M) with M_t the total mass of new entrants in period t. The law of motion for the distribution is:

$$\mu_{t+1} = \Gamma(\psi_t, \mu_t, M_{t+1}) \tag{3.14}$$

We define Γ below. The evolution of the distribution thus depends on bank entry and exit as well as on the decisions of the incumbent banks-varieties.

3.3.5 Credit Margins and Markups

Private bank-level *margins* $\chi(j)$ are defined as the ratio of capital relative prices $p(j)$ to the cost of funds $\bar{R}(j)$:

$$\chi(j) = \frac{p(j)}{\bar{R}(j)} \tag{3.15}$$

This is also precisely the definition of margins from our empirical analysis. Given the aggregate state vector $\mathbf{S}$ that we define below, the aggregate margin $X(\mathbf{S})$ is defined as the ratio of the aggregate price of capital to the average interest rate on deposits:

$$X(\mathbf{S}) = \frac{P(\mathbf{S})}{\bar{R}(\mathbf{S})} \tag{3.16}$$

In our model, $X(\mathbf{S})$ is a dynamic, endogenous object. That is, each $\chi(j)$ is determined conditional on the forward-looking expectations about the evolution of idiosyncratic return risk, the tightness of the leverage constraint, and the dynamic distribution of assets. The aggregate margin, in turn, is a time-varying average of the cross-section. Because of non-linearities and aggregate uncertainty, this average does *not* correspond necessarily to the margin of the average intermediary.

Although in order to compute margins in the full model we need to resort to numerical methods, we can derive several analytical results to show how credit margins are determined in a simpler version of the model. First, assume that we solve a static problem instead of the dynamic one. That is, ignore aggregate uncertainty. Second, treat v(j), $\bar{R}(j)$, and $R^T(j)$ as given. Third, assume the occasionally binding leverage constraint is always slack. We can solve for the bank-level rate-setting rule:

Proposition 4 (Bank Price-Setting Rule)**.** *The price-setting rule* $\frac{p(j)}{P(\boldsymbol{S})}$ *on each local credit market j is:*

$$\frac{p(j)}{P(\boldsymbol{S})} = \left[\frac{\beta\theta - 1}{\theta - 1} \frac{\bar{R}(j)}{\left(1 - \upsilon(j)\right) R^T(j)} K(\boldsymbol{S})^{\beta - 1}\right]^{\frac{1}{\theta(\beta-1)}} \tag{3.17}$$

and the marginal cost is:

$$MC(j) := \frac{\beta\theta - 1}{\theta} p(j) \frac{\bar{R}(j)}{\left(1 - \upsilon(j)\right) R^T(j)} \left[\left(\frac{p(j)}{P(\boldsymbol{S})}\right)^{-\theta} K(\boldsymbol{S})\right]^{\beta - 1} \tag{3.18}$$

Proof: Appendix 3.10

The proposition clarifies an important distinction between credit *margins* and *markups*. A credit margin, $\chi(j)$, is a measurable ratio of the credit rate p(j) over the deposit rate $\bar{R}(j)$. In our model, $\chi(j)$ is endogenous, heterogeneous across local credit markets, and aggregate state-dependent. The markup, however, given the CES assumption, is *constant* across all dimensions and depends only on the structural parameter θ.

In addition, the proposition clarifies the determinants of bank j's endogenous marginal cost. First, it is a function of the *relative* cost of funds $\frac{\bar{R}(j)}{R^T(j)}$. This is due to the fact that the marginal revenue, unlike standard models of monopolistic competition in product markets, depends not only on revenues p(j)k(j) but also on the return to investment $R^T(j)$. Second, there is a scale effect factor in $K(\mathbf{S})^{\beta-1}$. Finally, the marginal cost is negatively related to the probability of default $\nu(j)$.

3.3.6 The Incumbent Banker Problem

We now detail the dynamic problem of the incumbent intermediary. We follow recursive notation from now on: the solution does not depend on the specific local credit market j but only on the relevant state variables. Define $\mathbf{s} = \{n, \xi\}$ and $\mathbf{S} = \{\psi, \mu, M\}$ as the bankers' idiosyncratic and aggregate state vectors, respectively. Conditional on the state vector $\{\mathbf{s}, \mathbf{S}\}$, each banker maximizes its franchise value which is defined as the discounted stream of future flows of net worth. The bank discounts the future by adopting and augmenting the household's stochastic discount factor $\Lambda(\mathbf{S})$, which is determined in equilibrium together with household optimization. Each banker takes as given the aggregate uncertainty process $\{\psi\}$, aggregate quantities $\{\mathrm{K}(\mathbf{S}), \mathrm{N}(\mathbf{S})\}$, aggregate prices $\{\mathrm{P}(\mathbf{S}), \mathrm{R}^{\mathrm{k}}(\mathbf{S})\}$, variety-specific deposit rates $\bar{\mathrm{R}}(\mathbf{s}, \mathbf{S})$, and the law of motion of the distribution Γ. The generic incumbent banker therefore solves:

$$\mathrm{V}(\mathbf{s}, \mathbf{S}) = \max_{\{\mathrm{k,p,d}\}} \left\{ \mathbb{E}_{\mathbf{S}'|\mathbf{S}} \left[\Lambda(\mathbf{S}') \Big((1-\sigma)\mathrm{n}' + \sigma \mathrm{V}(\mathbf{s}', \mathbf{S}') \Big) \right] \right\} \tag{3.19}$$

s.t. conditions 3.1-3.4 and 3.6-3.14

We can simplify the problem substantially by rewriting it as a one-argument problem. Each bank now chooses the leverage ratio $\phi = \frac{\mathrm{pk}}{\mathrm{n}}$ to:

$$\max_{\phi} \left[\mu_{\mathrm{a}} \phi + \nu_{\mathrm{a}} \right] \tag{3.20}$$

subject to the same constraints as before, where $\mu_{\mathrm{a}} = (1-\upsilon)\tilde{\Lambda}(\mathbf{S}') \left[\mathrm{R}^{\mathrm{T}}(\mathbf{s}', \mathbf{S}') - \mathrm{k}^{\beta-1}\bar{\mathrm{R}}(\mathbf{s}, \mathbf{S}) \right]$ is the excess return on risky investments, and $\nu_{\mathrm{a}} = (1-\upsilon)\tilde{\Lambda}(\mathbf{S}')\bar{\mathrm{R}}(\mathbf{s}, \mathbf{S})$ is the cost of liabilities. In both instances, $\tilde{\Lambda}(\mathbf{S}') = \Lambda(\mathbf{S}') \left(1 - \sigma + \sigma \mathrm{V}(\mathbf{S}') \right)$ is the *augmented* stochastic discount factor. Note how the financial intermediary stochastic discount factor depends on the household's discount factor (because all bank equity ultimately belongs to the consumer) and on the exogenous exit rate σ.

The leverage ratio, when internalizing the credit demand system, can be written:

$$\phi = \frac{\mathrm{pk}}{\mathrm{n}} = \left(\frac{\mathrm{k}}{\mathrm{K}(\mathbf{S})} \right)^{-\frac{1}{\theta}} \frac{\mathrm{P}(\mathbf{S})\mathrm{k}}{\mathrm{n}} \tag{3.21}$$

From equation 3.21 we see how leverage depends on relative prices p(j), which in turn are set conditional on the elasticity of substitution θ. Varying the degree of banking competition feeds directly to the bank's preference for risk-taking.

For illustration, we now contrast our ϕ with the standard definitions of market leverage from GK models. To get to the GK version precisely, we need to set $\theta \to \infty$ and $\beta = 1$, as well as $\sigma_{\xi} = 0$. Bank heterogeneity then collapses to the case of a representative, homogeneous intermediary. In that case, bank leverage becomes $\phi_{\mathrm{GK}} = \frac{\mathrm{P}(\mathbf{S})\mathrm{k}}{\mathrm{n}}$, which can be seen cleanly from Equation 3.21. Because all relative prices collapse to unity, the market power channel is shut off.

Additionally, it can be seen from the expression for μ_{a} above how the value function depends on bank characteristics when $\beta > 1$. Notice how μ_{a} depends explicitly on the choice of k. With decreasing returns, the bank internalizes the impact of balance sheet choices on excess returns which in turn feed into the value of the franchise. With constant returns, bank-level characteristics do not matter and the state vector includes only aggregate variables. We discuss the role of bank size heterogeneity more formally in the next section.

3.3.7 Bank Size Heterogeneity and Scale Variance

An important departure of our framework from Gertler and Kiyotaki (2010) and Gertler and Karadi (2011) is that the value function of the incumbent is *not* linear in bank-level net worth. That is, the solution is not invariant to the scale of the intermediary. The role of bank size heterogeneity can be best visualized from the following proposition, where we re-introduce the (j) notation for additional emphasis on heterogeneity:

Proposition 5. *The solution to the incumbent banker's problem for each j, conditional on* $\beta > 1$*, the aggregate state vector* $\boldsymbol{S}$*, initial net worth* $\mathrm{n(j)}$ *and idiosyncratic state* $\xi\mathrm{(j)}$ *is*

$$\mathrm{V}\Big(\mathrm{n(j)}, \xi\mathrm{(j)}; \boldsymbol{S}\Big) = \zeta\Big(\mathrm{n(j)}, \xi\mathrm{(j)}; \boldsymbol{S}\Big)\mathrm{n(j)}$$

where the marginal value of net worth is:

$$\zeta\Big(\mathrm{n(j)}, \xi\mathrm{(j)}; \boldsymbol{S}\Big) = \frac{\Big(1 - \upsilon\mathrm{(j)}\Big)\mathbb{E}\left(\Lambda(\boldsymbol{S}')\Big[1 - \sigma + \sigma\zeta\Big(\mathrm{n}'\mathrm{(j)}, \xi'\mathrm{(j)}; \boldsymbol{S}'\Big)\Big]\mathrm{k(j)}^{\beta-1}\bar{\mathrm{R}}\mathrm{(j)}\right)}{1 - \varphi\Big(\mathrm{n(j)}, \xi\mathrm{(j)}; \boldsymbol{S}\Big)} \tag{3.22}$$

and the multiplier on the moral hazard leverage constraint is

$$\varphi\Big(\mathrm{n(j)}, \xi\mathrm{(j)}; \boldsymbol{S}\Big) = \max\left[1 - \frac{\Big(1 - \upsilon\mathrm{(j)}\Big)\mathbb{E}\left(\Lambda(\boldsymbol{S}')\Big[1 - \sigma + \sigma\zeta\Big(\mathrm{n}'\mathrm{(j)}, \xi'\mathrm{(j)}; \boldsymbol{S}\Big)\Big]\mathrm{k(j)}^{\beta-1}\bar{\mathrm{R}}\mathrm{(j)}\right)\mathrm{n(j)}}{\lambda\mathrm{k(j)}^{1-\frac{1}{\theta}}\Big(\mathrm{K}(\boldsymbol{S})\Big)^{\frac{1}{\theta}}\mathrm{P}(\boldsymbol{S})}, 0\right] \tag{3.23}$$

Proof: Appendix 3.10

Proposition 5 shows that bank-level characteristics matter for aggregation. In particular, the distributions of both bank net worth *and* idiosyncratic return shocks $\xi\mathrm{(j)}$ are state variables. The former is due to $\beta > 1$ and $\theta > 1$, whereas the latter is due to the persistence of the shock process and the incomplete markets assumption.

3.3.8 Cyclical Entry and Exit

There is a large number of new, potential financial varieties which are managed by financiers. Before entry, the financier pays a fixed equity issuance cost e in units of aggregate capital. After paying the sunk cost, the financier receives an idiosyncratic return profitability draw $\xi_0 \in \Xi$ from the ergodic distribution $\mathrm{G}_0(\xi)$ that governs ξ. The new entrant is also bestowed with a starting level of net worth $n_0(\mathbf{S}) = \iota N_t$ where $0 < \iota < 1$. This assumption is motivated by pro-cyclical bank entry in the data. Immediately afterwards, the entrant decides whether to operate or to exit. Conditional on staying, the financier becomes a new financial variety and adds to J_t. Conditional on its startup state vector $\{n_0(\mathbf{S}), \xi_0\}$, the entrant operates if its expected discounted franchise value exceeds e. The value function for the entering financial variety is therefore:

$$V^e(n_0, \xi_0; \mathbf{S}) \equiv \max\left[\mathbb{E}V(\mathbf{s}; \mathbf{S}) - e, 0\right] \tag{3.24}$$

In equilibrium, either V^e is equal to 0, the number of entrants is 0, or both. The incumbents are subject to two sources of exit risk: the involuntary exogenous exit rate σ and the endogenous probability of default $\nu(j)$ which is specific to every local market. If a financial variety exits, its market niche would never be taken over by any of the incumbents. Importantly, the extensive margin dynamic depends explicitly on the state of the business cycle. The two main factors that impact financiers' entry decisions are bank franchise values $V(\mathbf{S})$ and startup equity injections $n_0(\mathbf{S})$. Both are procyclical since aggregate demand for credit (and the stock of capital) are positively related to ψ. That immediately implies that equilibrium entry is also procyclical, in line with the data.

3.3.9 Dynamics of the Cross-Sectional Distribution of Banks

Define π^e as a mass of financial varieties that exits either due to endogenous default or because of the exogenous exit shock. The cross-sectional distribution evolves according to:

$$\mu'(n, \xi_i) = (1 - \pi'^e) \int_{\{(n,\xi_i)|K(n,\xi_i;\mathbf{S})\in\mathbf{B}\}} G_{ji}\mu(dn, d\xi_i) + M' \int_{\{(n_0,\xi_i)|K(n,\xi_i;\mathbf{S})\in\mathbf{B}\}} G_0(\xi_i), \forall \xi_i \in \Xi \tag{3.25}$$

Recall that $G_0(.)$ is the CDF of ξ for new entrants and $G_{x'x}$ is the Markov chain for ξ of the incumbent.

3.3.10 Households

For simplicity, assume inelastic labor supply normalized to 1. The representative household chooses the supply of deposits to each financial variety, $b_t(j)$, and consumption C_t, subject to the budget constraint:

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

subject to

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

where W_t is the equilibrium wage and π are profits (net of any transfers) from bank ownership redistributed back to the household lump sum. The first order condition for deposits yields the following equation:

$$\bar{R}_t(j) = \frac{1 - \upsilon_t(j)x_t(j)\mathbb{E}\left(R^T_{t+1}(j)\Lambda_{t+1}\right)}{\left(1 - \upsilon_t(j)\right)\mathbb{E}\left(\Lambda_{t+1}\right)} \tag{3.26}$$

where $\Lambda_{t+1} = \beta_h \frac{u'(c_{t+1})}{u'(c_t)}$ is the stochastic discount factor. Deposits are risky because there is no deposit insurance. The consumer prices bank default risk into the distribution of variety-specific deposit rates, which depend on the deposit *recovery rate* $x_t(j)$:

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

with $\phi_t(j)$ being the market leverage ratio, defined as before.

3.3.11 Recursive Industry Equilibrium

Credit market clearing requires:

$$K(\mu, \mathbf{S}) = \int_{\mathbf{B}} \left(k(\mathbf{s}; \mathbf{S})\right)\mu(dn, d\xi) + M(\mathbf{S})\int_{\mathbf{B}} \left(k(n_0, \xi_0; \mathbf{S})\right)dG(\xi_0) + M(\mathbf{S})e \tag{3.27}$$

where the first term on the right hand side is total demand by incumbents, the second term is total demand by entrants, and the final term is the entry cost paid by the entrants. Similarly, equilibrium in the market for bank deposits requires:

$$\int_0^J b(j)dj = \int_{\mathbf{B}} \left(d(\mathbf{s}; \mathbf{S})\right)\mu(dn, d\xi) \tag{3.28}$$

Goods market clearing requires production goods to be used either for household consumption or firm investment. The latter includes investment demand that is intermediated both by the incumbent and the new entrants.

$$Y(\mathbf{S}) = C(\mathbf{S}) + I(\mathbf{S})$$

A *recursive industry equilibrium* is defined as a set of functions that include the value function of the banker $V(\mathbf{s}; \mathbf{S})$, optimal policies for bank capital investment $k(\mathbf{s}; \mathbf{S})$ and deposit demand $d(\mathbf{s}; \mathbf{S})$, household consumption $c(b_{-1}; \mathbf{S})$ and deposit supply $b(b_{-1}(j); \mathbf{S})$, the mass of new bank entrants $M(\mathbf{S})$, competitive wage $W(\mathbf{S})$ and capital $R_k(\mathbf{S})$ pricing functions, the aggregate price index of financial varieties $P(\mathbf{S})$, a marginal utility process $\Lambda(\mathbf{S})$, and the menu of market-clearing deposit rates $\bar{R}(\mathbf{s}; \mathbf{S})$ such that:

1. The household's choices $\{C(\mathbf{S}), b(j)\}$ are optimal conditional on $\{W(\mathbf{S}), \bar{R}(\mathbf{s}; \mathbf{S})\}$.
2. The banker's choices $\{k(\mathbf{s}; \mathbf{S}), p(\mathbf{s}; \mathbf{S}), d(\mathbf{s}; \mathbf{S})\}$ are optimal conditional on $\{\psi, \Lambda(\mathbf{S}), K(\mathbf{S}), P(\mathbf{S}), \bar{R}(\mathbf{s}; \mathbf{S}), \mu(\mathbf{S})\}$
3. The returns on factors of production are: $R^k(\mathbf{S}') = \psi' \left(\frac{\alpha AK(\mathbf{S}')^{\alpha-1} + (1-\delta)P(\mathbf{S}')}{P(\mathbf{S})} \right)$, $W(\mathbf{S}) = (1-\alpha)K(\mathbf{S})^\alpha$.
4. $\{K(\mathbf{S}), D(\mathbf{S}), N(\mathbf{S})\}$ are consistent with the cross-sectional distribution and the monopolistic credit demand system in (3.4)-(3.7).
5. The free-entry condition (3.24) is satisfied and is consistent with individual choices.
6. The credit market clears as in (3.27). The deposit market clears for each variety j as in (3.28).
7. The cross-sectional distribution evolves according to (3.25) and is consistent with bank-level demand functions.

3.3.12 Symmetric, Non-Stochastic Steady State

To shed more light on the equilibrium properties of the model, it is useful to inspect features of the symmetric, non-stochastic steady state. In doing so, and following the result in Proposition 4, we pin down the aggregate interest rate on deposits. It is determined in a static equilibrium where aggregate price levels and interest rates are obtained simultaneously for a given level of aggregate demand for credit. To that goal, we require additional assumptions, for tractability. First, we set $\xi(j)$ to the ergodic mean for all j. Second, we consider symmetric equilibria only, i.e., when $p(j) = P\ \forall j$. By extension, this implies homogeneous probabilities of default $\upsilon(j)$ and deposit interest rates. Note that this corresponds to analyzing a representative (average) banker from the distribution, rather than shutting down the default risk or idiosyncratic returns channels completely. The aggregate cost of funds is determined in the following proposition.

Proposition 6 (Aggregate Rate Rule)**.** *The aggregate rate-setting rule in the banking sector is*

$$\bar{R} = \left[\frac{\beta\theta - 1}{\theta - 1} \frac{K^{\beta-1}}{R^T(1-\nu)} \right]^{-1} \tag{3.29}$$

Proof. Follows immediately from Proposition 4 after assuming symmetry in the price-setting rule, i.e. in equation 3.17 set $p(j) = P\ \forall j$ and $R(j) = R\ \forall j$.

□

The aggregate rate-setting schedule is a downward-sloped demand curve for bank lending. Note that the slope of the line is independent of the elasticity of substitution, which acts as a horizontal curve shifter. In the limiting case of θ approaching infinity, the special case of perfect competition in the credit market is achieved. A symmetric equilibrium is defined by the intersection with the upward-sloping funding cost rule, equation 3.26 from the household's problem. It is straightforward to see that the rate is increasing in K because (a) $\bar{R}(j)$ is decreasing in the deposit recovery rate x(j) and (b) x(j) is decreasing in the leverage ratio $\phi(j)$ and k(j). Everything else equal, as debt-financed capital grows, the recovery rate falls and the deposit rate goes up. In the symmetric equilibrium, the average rate $\bar{R}$ is increasing in the aggregate stock of capital K.

A static, symmetric equilibrium generally exists when $\beta > 1$. We can visualize the equilibrium graphically, for some given values of ν and $R^T > 0$ as well as $\beta > 1$ and $\theta > 1$. Figure 3.6 portrays graphically symmetric equilibria under the alternative scenarios of perfect (PC) and monopolistic (MC) competition in lending. Monopolistic competition is our baseline case when $\theta > 1$ but is finite. Perfect competition is the theoretical limit of $\theta \to \infty$. The downward-sloping curve is the aggregate rate-setting rule (Equation 3.29). On the horizontal axis we have quantities - the aggregate demand for credit K. On the vertical axis are aggregate interest rates.

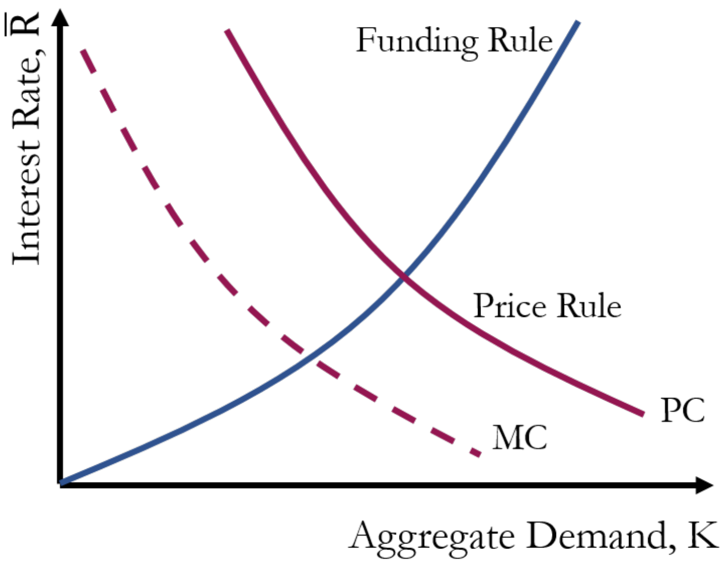
As θ falls, greater credit market power shifts the pricing rule leftward with no immediate effects on the funding schedule. The wedge between the MC and PC curves grows and the social deadweight loss from credit markups increases. This yields a decline in both aggregate demand and the cost of bank funds. Credit market power leads to an *aggregate under-lending externality*. In the monopolistic competition (MC) equilibrium, aggregate demand is strictly lower than in the perfect competition (PC) counterfactual.[8] Our result parallels the canonical under-utilization of labor resources in Blanchard and Kiyotaki (1987). Note that the only resource used in the production of financial varieties is capital k(j), thus the under-utilization of risky capital in equilibrium. In a more sophisticated setup, if the banker also employed labor in order to supply each additional unit of k(j), labor would be also underutilized in equilibrium. Assuming no market power on the liability side of the balance sheet, the credit supply externality would only strengthen in this case.

3.3.13 Numerical Algorithm

There are several considerable computational challenges in solving our model. First, the financial sector needs to construct forecasts for the return to aggregate capital $R^{k'}$, which depends on the forecasts for K' and P'. Both aggregates, in turn, potentially depend on the whole distribution

[8]This result stems from our assumption that all non-financial firms depend on bank funding. In practice, the relevance of the externality scales with the share of bank-dependent firms in the distribution. Clearly, if firms can finance investment with retained earnings or equity issuance, the credit supply externality would have little to no bite. This observation is true not only for our environment but also for the general class of GK models.

Figure 3.6: **Monopolistic Banking Equilibrium - Visualization**



Notes: This figure visualizes the static, symmetric equilibrium with monopolistic competition in bank lending. The downward-sloping curve is the aggregate pricing rule from Equation 3.29. The upward-sloping line is the deposit supply schedule from the household's problem. The graph highlights equilibria under monopolistic (MC) and perfect competition (PC) in the bank credit market. The MC case reflects the baseline scenario of a finite $\theta > 1$. The PC equilibrium is approximated by $\theta \rightarrow \infty$. On the horizontal axis is aggregate demand for credit K. On the vertical axis is the aggregate rate $\bar{R}$.

of bank assets k(j) - an infinitely dimensional object. We solve this problem with a variant of the stochastic simulation and partial aggregation method of Maliar et al. (2010). This method builds on the general algorithm for solving models with heterogeneous agents and aggregate uncertainty that was originally proposed in Krusell and Smith (1998). In the algorithm, linear forecasts are cast on a number of *moments* of the banking distribution. Specifically, we assume a linear projection of K' and P' on m moments of the k(j) distribution. In the baseline case, we will track only the mean. In Section 3.6.4 we will also track the second moment. Second, the deposit market must clear at all points of the state space. Third, the bank adopts and augments the households' stochastic discount factor $\Lambda(S)$, which is an endogenous state variable that must be kept track of. Fourth, the dynamic distribution of bank net worth must be consistent with both endogenous entry and exit. Finally, the bank faces an occasionally binding leverage constraint which could bind on any part of the idiosyncratic or aggregate state space. Below we describe all steps of our algorithm in more detail.

0. Solve a simpler version of the model without aggregate uncertainty. Use the solution as an initial condition for the full model with aggregate risk. Construct grids for aggregate capital and the aggregate shock. For individual bank net worth and household wealth, we use an unequally spaced grid with more points on lower values of n(j) and b_{-1}.

1. Solve the problem of the representative household given equilibrium wages and deposit rates. We use the endogenous gridpoints method for speed (Carroll, 2006). Construct $\Lambda(\mathbf{S})$.

2. Solve the financial intermediation problem in three steps

   (a) Conjecture a starting law of motion for aggregate capital $\log(\mathrm{K}') = \eta_{\mathrm{k}}\Big(\log(\mathrm{m}_{\mathrm{k}})\Big)$ and the projection for prices $\log(\mathrm{P}') = \eta_{\mathrm{p}}\Big(\log(\mathrm{m}_{\mathrm{k}})\Big)$. Construct $\mathrm{R}^{\mathrm{k}'}$.

   (b) Given η_{k}, η_{p}, Λ and initial guesses for $\mathrm{V}(\mathrm{j})$, $\nu(\mathrm{j})$, and $\bar{\mathrm{R}}(\mathrm{j})$ solve the bankers' problem using value function iteration. To handle the occasionally binding leverage constraint, first guess that the constraint is binding on a point in the grid space. Compute the implied Lagrange multiplier. If the constraint is slack, re-solve the problem using global maximization routines while ignoring the constraint.

   (c) Compute the implied policy function for deposit demand. Update the interest rates on deposits, compare with the initial guess, and iterate upon convergence is achieved on each point of the state space.

3. Using the newly computed policy functions, run a large simulation of N varieties over T quarters where incumbents are subject to both idiosyncratic and aggregate shocks. Compute the implied time-varying distribution of bank net worth n(j), the supply of investment into firm claims k(j), and the prices of claims p(j).

4. Solve the optimal entry problem in 3.24 and determine the time-varying mass of entrants M for each quarter of T.

5. Construct the sequence of the aggregate demand for capital K that tracks demand of the incumbent and of the new entrants. Run linear regressions $\mathrm{K}' = \hat{\eta}_{\mathrm{k}}(\mathrm{m}_{\mathrm{k}})$ and $\mathrm{P}' = \hat{\eta}_{\mathrm{p}}(\mathrm{m}_{\mathrm{k}})$. Compare regression coefficients with the original η_{k} and η_{p}. If convergence is not achieved, return to step 2(a) and continue until convergence of the inner loop.

6. When the inner loop converges, construct updated versions of competitive wages and deposit rates using optimal $\nu(\mathrm{j})$ and $\phi(\mathrm{j})$. Return to step 1, re-solve the household problem and construct a new $\Lambda(\mathbf{S})$. Continue with the inner loop. Keep iterating the program until convergence of the outer loop is achieved. Accuracy of the algorithm is discussed in Appendix 3.11.

3.4 Quantifying the Model

We begin a quantitative illustration of the model by first reporting the details of our calibration.

3.4.1 Calibration

Table 3.3 displays the parameter values chosen for the model calibration. The unit of time in the model is a quarter. We start by describing standard macro parameters. The share of aggregate capital in production is set to $\alpha = 0.36$. The discount rate β_{h} =0.996 targets the steady-state annual risk-free rate of 1.56%. Aggregate capital $\mathrm{K_t}$ depreciates at the quarterly rate $\delta = 0.025$. We assume log-utility in consumption ($\sigma_{\mathrm{h}} = 1$).

For parameters in the banking block, the dividend payout ratio is set to $\sigma = 0.9$. This number is broadly consistent with the exit rate of financial intermediaries in the U.S. According to the Federal Deposit Insurance Corporation, there were about 11000 commercial banks in the U.S. at the start of 1980. This number has fallen below 5000 by 2020. This implies an approximate annual exit rate of 3% and a life expectancy of a banker of about 8.25 years. In the model, that number is 10 conditional on zero default risk. The fraction of bank assets that are divertible by the manager is $\lambda = 0.12$. This number targets a steady state bank leverage ratio of 8. Endowment of new entrants and the fixed cost of entry are calibrated in order to keep the net entry rate relatively stable over the cycle at around 5%.

The monopolistic banking block requires calibration of two main parameters. First, the elasticity of substitution $\theta = 2$ implies an average stationary credit margin of 1.48 over the cost of funds, which is broadly in line with the existing evidence on loan margins in the financial sector (De Loecker et al., 2020; Diez et al., 2018). Jamilov (2020) estimates the average nationwide elasticity of substitution across U.S. commercial banks to be roughly 1.2, and we pursue a more conservative calibration approach. The returns to scale parameter of $\beta = 1.01$ suggests minor dis-economies of scale and almost-constant returns. The calibrated value is consistent with empirical evidence on some dis-economies of scale in European banking (Anolli et al., 2015).

Calibration of the idiosyncratic return process follows closely the recent work by Galaasen et al. (2020). Galaasen et al. (2020) estimate the pass-through of idiosyncratic firm shocks on bank-level returns using matched bank-firm data from Norway. They estimate the shock process and find annual persistence and standard deviation parameters of $\rho_{\xi} = 0.11$ $\sigma_{\xi} = 0.25$. That is, the idiosyncratic rate of return shock in banking is (a) volatile and (b) not very persistent. These values are in line with our chosen quarterly parameters in the Table. The fraction of financial wealth exposed to idiosyncratic risk $\kappa = 0.5$ is in line but slightly on the upper side of the pre-financial crisis share of the shadow banking business in overall U.S. banking (Gorton and Metrick, 2010). For the aggregate shock process, we pick σ_{ψ} and ρ_{ψ} such that output volatility in the model corresponds roughly to that of the pre-Crisis period. Both idiosyncratic and aggregate processes, in order to keep the state space manageable, are discretized with the Tauchen (1986) method.

Table 3.3: **Model Parameters**

Parameter	Description	Value
	Standard Macro	
α	Share of capital in production	0.36
β_h	Discount factor	0.996
σ_h	Household risk aversion	1
δ	Capital depreciation rate	0.025
	Banking	
σ	Dividend payout ratio	0.9
λ	Share of divertible assets	0.12
ι	Entry starting endowment	30% of N_t
e	Entry fixed cost	0.11
	Monopolistic Competition	
θ	Credit demand elasticity	2
β	Local returns to scale	1.01
	Idiosyncratic and Aggregate Risk	
κ	Fraction of returns exposed to idiosyncratic risk	0.5
ρ_ξ	Serial correlation of idiosyncratic risk	0.5
σ_ξ	SD of idiosyncratic risk	0.1
ρ_ψ	Serial correlation of aggregate risk	0.914
σ_ψ	SD of aggregate risk	0.015

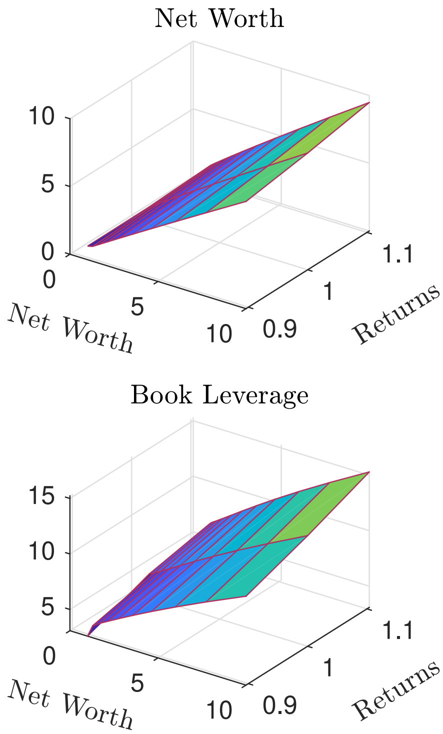
Notes: Parameters that are exogenously fixed in the model calibration.

3.4.2 Model Policy Functions

Figures 3.7 and 3.8 plot two-dimensional policy functions for key financial-sector quantities, prices, and measures of risk. The two dimensions are the idiosyncratic states: initial net worth (size) $n_t(j)$ and returns $\xi_t(j)$. Figure 3.7 reports quantities. (Book) leverage is defined as total assets over total net worth. Overall, banks with larger $n_t(j)$ and $\xi_t(j)$ tend to choose greater quantities of future net worth $n_{t+1}(j)$ and assets $k_{t+1}(j)$. Such banks are also more (book) levered, which is consistent with the empirical facts from Coimbra and Rey (2019) and Gopinath et al. (2017) on both financial and non-financial firms. This is also in line with our empirical analysis from Section 3.2.2. Finally, larger and more profitable banks have a greater franchise value $V_{t+1}(j)$.

Figure 3.8 reports policy functions of prices and measures of risk. Generally, credit margins, $\chi(j)$, decline with bank size and profitability in the cross-section. This feature of the model is a consequence of us working with CES aggregation. Noticeably, it is consistent with the empirical

Figure 3.7: **Model Policy Functions - Quantities**



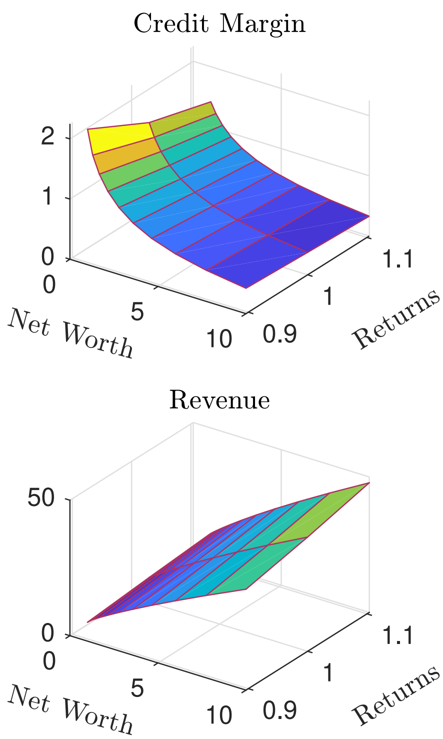
Notes: Optimal choices represented as two-dimensional surfaces. The two dimensions are idiosyncratic state variables in the banking sector: initial net worth n(j) and idiosyncratic return ξ(j). Leverage is in book values.

evidence in Section 3.2.2 documenting that banks with higher net worth charge lower margins. Larger, more profitable banks also earn more in total revenues $p_t(j)k_t(j)$. Small institutions face elevated risks of insolvency $\nu_t(j)$, which is priced in the interest rate on deposits $\bar{R}(j)$.

3.4.3 Ergodic Cross-Sectional Distributions

Figures 3.9 and 3.10 present two-dimensional histograms for bank book leverage, default risk, relative prices, and deposit rates. These cross-sectional distributions are ergodic, i.e., obtained from the recursive general equilibrium with aggregate uncertainty. One of the two axes represents net worth (size) n(j). The other axis is the plot-specific variable of interest. The distribution of leverage is centered around 8, our steady state target. It is right-skewed, i.e., there is a small number of banks with high leverage ratios. This is consistent with the cross-sectional facts reported in Section 3.2.2 above. Leverage also grows with initial net worth, which was also visible from the model policy functions. The distribution of margins χ(j) is centered around 1.5, consistent with our target. Credit margins decline with the size of the intermediary, as mentioned before. Finally, the distribution of

Figure 3.8: **Model Policy Functions - Prices**



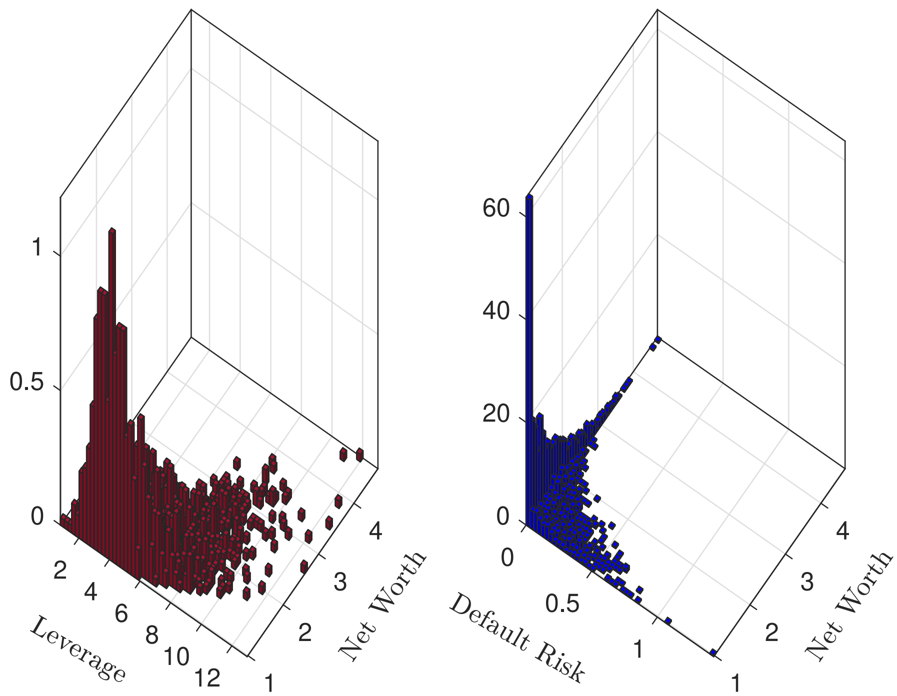
Notes: Optimal choices represented as two-dimensional surfaces. The two dimensions are idiosyncratic state variables in the banking sector: initial net worth n(j) and idiosyncratic return ξ(j). Deposit rates are in percent, annualized. Default risk is in percent, annualized.

default (insolvency) risk is concentrated in the left tail of the bank net worth distribution. Small institutions face a probability of insolvency that is as high as 1% on a yearly basis. This distribution is priced into the cross-section of deposit rates $\bar{\mathrm{R}}$(j): smaller, riskier institutions must compensate the risk-averse household for "investing" into a risky bank. In equilibrium, the premium is reflected in higher interest rates on deposits.

3.4.4 Credit Cycle Statistics

Table 3.4 reports unconditional standard deviations and correlations with output from our model economy. We focus on the financial cycle and the first three moments of the distributions of bank assets, net worth, credit margins, default risk, and book leverage. In order to obtain these moments, we simulate the model for 10,000 quarters. We see that model-based standard deviations are smaller than in the data. This is due to the fact that our statistics are based on a model with a single aggregate shock ψ. We correctly match the regularity that higher-order moments of the distributions (especially concentration) are typically more volatile than the mean.

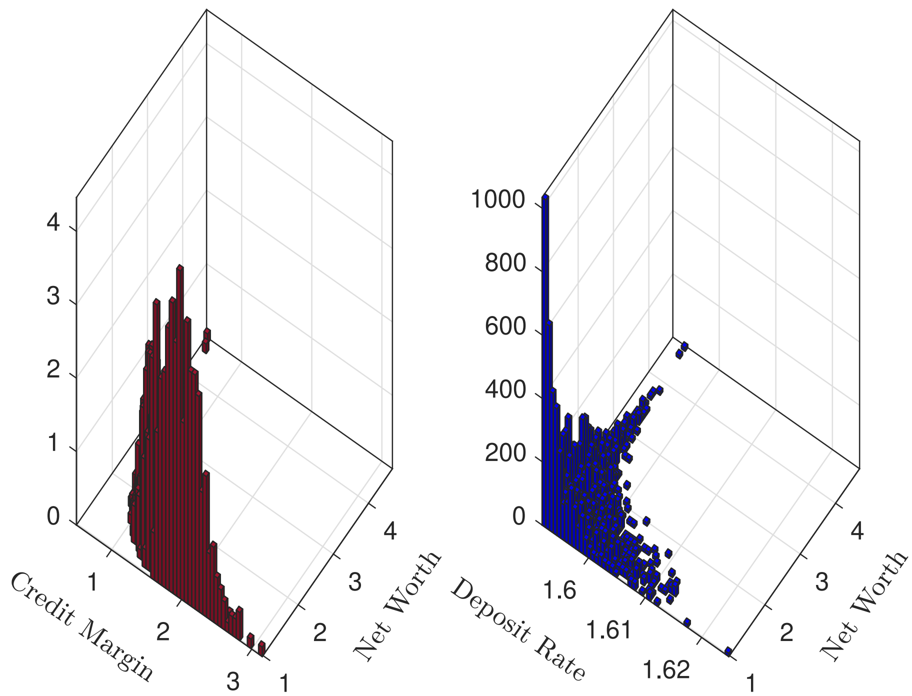
Figure 3.9: **Ergodic Distributions - Risk-Taking**



Notes: ergodic cross-sectional distributions obtained from the recursive competitive equilibrium. Default risk is in percent, annualized. Leverage is in book values. Figures plot the pdf normalization.

The table also reports model-implied credit cycle correlations. In terms of correlations, our model can match the data well. We match the pro-cyclicality of the first two moments of financial assets and net worth. We also obtain - correctly - negative correlations for the concentration of both assets and net worth. In the model credit margins are counter-cyclical, which is in line with the data. The model predicts pro-cyclical second and third moments of the margins distribution, while in the data those moments are negative and positive, respectively. All three moments of the distribution of default risk have correct cyclical properties. Book leverage in the model is pro-cyclical in all three moments. In the data, concentration of leverage is counter-cyclical. Finally, both the total number of incumbent intermediaries and the mass of new entrants is correctly pro-cyclical in the model. Overall, the model is able to replicate 15 out of the 17 free, untargeted correlations, including the entry rate and the number of active intermediaries, which are both pro-cyclical as also reported in Corbae and D'Erasmo (2019).

Figure 3.10: **Ergodic Distributions - Prices**



Notes: ergodic cross-sectional distributions obtained from the recursive competitive equilibrium. Deposit rates are in percent, annualized. Figures plot the pdf normalization.

3.4.5 Financial Recessions

Next we characterize equilibrium dynamics in the model. We study the model behavior in response to a (one standard deviation) negative ψ (capital quality) shock. This shock is representative of a worsening of banks' balance sheets. Our numerical approach consists of two general steps. First, we run a "benchmark" simulation of our model economy for 10,000 periods. Second, in some quarter T*, the economy is hit with a counterfactually low realization of the aggregate capital quality shock ψ_{T*}. It is then allowed to revert back to the benchmark path with its normal autocorrelation coefficient while being subjected to normal shocks as in the benchmark. The differential between the benchmark and the counterfactual simulations identifies the impact of the aggregate shock on each variable relative to its stochastic steady state value.

Figure 3.11 displays the responses of key macroeconomic aggregates. The crisis episode is associated with a contraction in output, consumption, and aggregate capital. The risk premium, defined in the model as $R^T(\mathbf{S}) - \bar{R}$, spikes upwards due to the deterioration of conditions in the financial sector: the value of productive capital, and by extension, bank assets has declined. This

Table 3.4: **Credit Cycle Statistics - Data and Model**

	Data		Model	
Variable	Standard Deviation	Correlation with Y_t	Standard Deviation	Correlation with Y_t
Assets Mean	13.383	0.498	2.316	0.798
Assets Dispersion	19.371	0.642	4.276	0.541
Assets HHI	18.281	-0.568	13.897	-0.103
Net Worth Mean	11.268	0.211	1.920	0.842
Net Worth Dispersion	18.076	0.544	3.837	0.683
Net Worth Concentration	16.64	-0.472	8.553	-0.238
Margins Mean	31.046	-0.563	0.765	-0.305
Margins Dispersion	42.404	-0.370	1.829	0.437
Margins Concentration	56.595	0.725	10.996	0.476
Default Mean	57.751	-0.325	6.887	-0.740
Default Dispersion	58.498	-0.309	6.040	-0.493
Default Concentration	32.021	0.033	21.058	0.278
Book Leverage Mean	6.036	0.701	0.679	0.197
Book Leverage Dispersion	6.855	0.043	1.817	0.097
Book Leverage Concentration	20.157	-0.641	16.937	0.043
Bank Entry Mass		0.700	0.195	0.810
Number of Banks			0.102	0.811

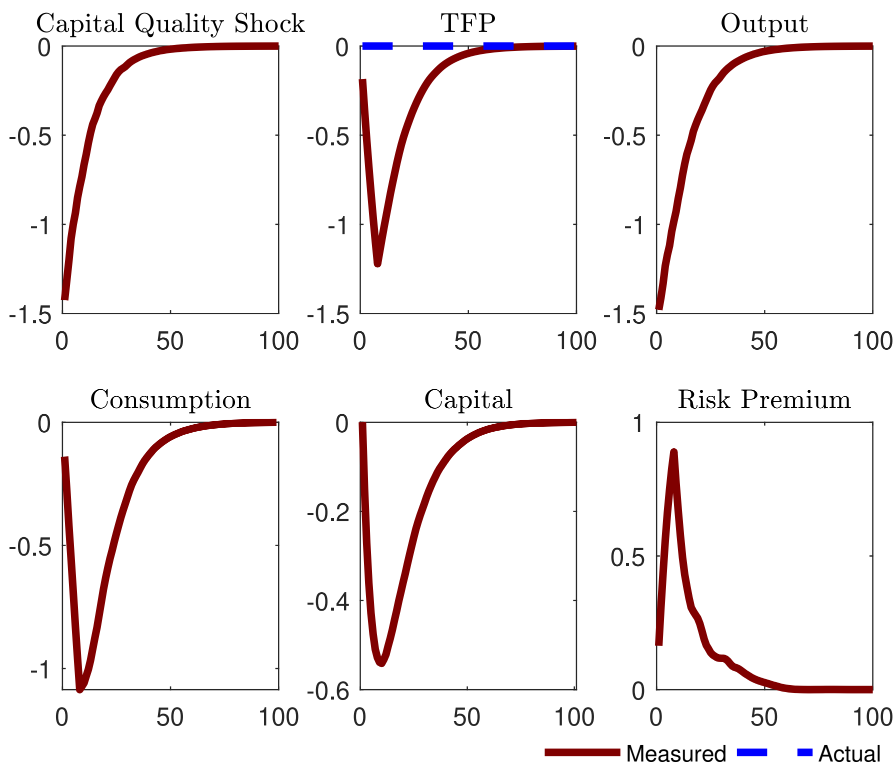
Notes: Table reports standard deviations and correlations with output of key financial aggregates. Columns (2-3) report moments in the data. Columns (4-5) report moments from the model. To obtain model-based moments we solve and simulate the model for 10,000 periods. All simulations are conditional on capital quality shocks only.

effect gets amplified through the tightening of bank leverage constraints and a further fall in asset values. The impact on the real economy runs directly through firms' reliance on bank financing and the collapse of non-financial investment. There is no aggregate uncertainty in total factor productivity (TFP) in the model, and so the actual A_t is unaffected. Measured TFP ($\tilde{A}_t$), however, defined as the simple Solow residual of the production function, declines in line with the rest of the real aggregates.[9]

Figure 3.12 presents the response of the financial sector. For each financial characteristic, we

[9]Formally, we define $\tilde{A}_t$ as the difference between the time-series of output and capital. However, Y_t is the actual, realized series while capital is from the stochastic steady state: $\tilde{A}_t = Y_t - K_t^{ss}$. That is, in response to unexpected ψ_t shocks, the econometrician can always measure and observe that Y_t is falling but cannot correctly attribute it to the "true" capital stock that is subject to financial frictions and imperfect competition. With measured capital not showing any response to the shocks, the econometrician attributes the unexplained component of the drop in Y_t to the measure of our ignorance.

Figure 3.11: **Financial Recession - Macro Outcomes**

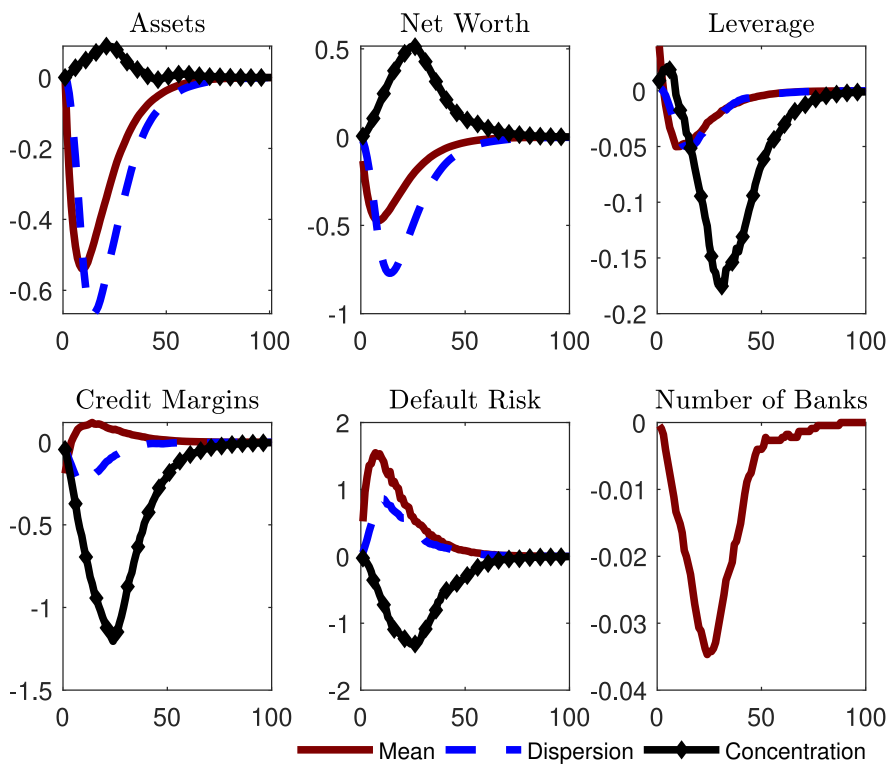


Notes: Response to a one standard deviation negative ψ_t shock. The benchmark economy is in the stochastic steady state. All figures plot percentage point differences. See main text for details.

plot responses of the mean, dispersion, and concentration of the corresponding cross-sectional distribution. First, the mean and dispersion of bank assets and net worth fall, while concentration levels of both rise. The negative aggregate shock shifts the distribution of bank size *leftward*. Concentration rises due to the uneven distributional effects in the financial sector. For one, bank exit due to default accelerates and bank size gets clustered around zero due to limited liability. On the other hand, institutions with ex-ante high net worth and returns are affected marginally less and become larger in relative terms than prior to the shock. Effectively, a form of *reallocation of credit provision* takes place that favors banks with ex-ante strong balance sheets and high profitability. Additionally, because startup equity of new entrants is tied up to the average level of net worth, all entrants begin with a lower level of net worth, which further boosts the size concentration.

Mean leverage (in book values) falls, and so do its dispersion and skewness. In the model book leverage is positively correlated with net worth. As the distribution of net worth shifts to the left, smaller intermediaries become less risky in terms of book leverage. New entrants begin with lower levels of initial equity and, conditional on operation, immediately choose very low levels of book leverage. This gets reflected in a reduced skewness of leverage.

Figure 3.12: **Financial Recession - Banking Outcomes**



Notes: Response to a one standard deviation negative ψ_t shock. The benchmark economy is in the stochastic steady state. All figures plot percentage point differences. See main text for details.

Furthermore, average credit margins rise, while their dispersion and skewness both fall. As seen from our model policy functions, credit margins decline with bank net worth. As the average intermediary becomes smaller, the average loan margin increases. However, the presence of a small fraction of very large intermediaries is reflected in the decline of the skewness of margins, in line with the data showing that the concentration of credit margins is pro-cyclical. A similar response pattern is observed for default risk: smaller institutions are considerably more likely to be driven to insolvency due to negative realizations of idiosyncratic return shocks. Large intermediaries with a sufficient buffer stock of net worth are immune to this risk, which makes the default risk distribution left-skewed.

Finally, the total number of banks (financial varieties) falls. This is the result of two drivers of the extensive margin. First, bank entry stalls because the cutoff level of initial profitability has increased. This is due to the fact that startup equity injections, which are tied to aggregate net worth, are now lower. Second, the bank exit rate, driven by endogenous default, rises. As a result, the number of active incumbent intermediaries shrinks.

Overall, and conditional on capital quality shocks only, the model is capable of reproducing the cyclical properties of all higher-order moments except for two (skewness of leverage and dispersion of margins).

3.5 Inspecting the Model Mechanisms

In this section we build on the "financial recession" experiment of the previous section and isolate the contributing roles of the three key model mechanisms: (i) bank market power, (ii) idiosyncratic risk, and (iii) endogenous entry.

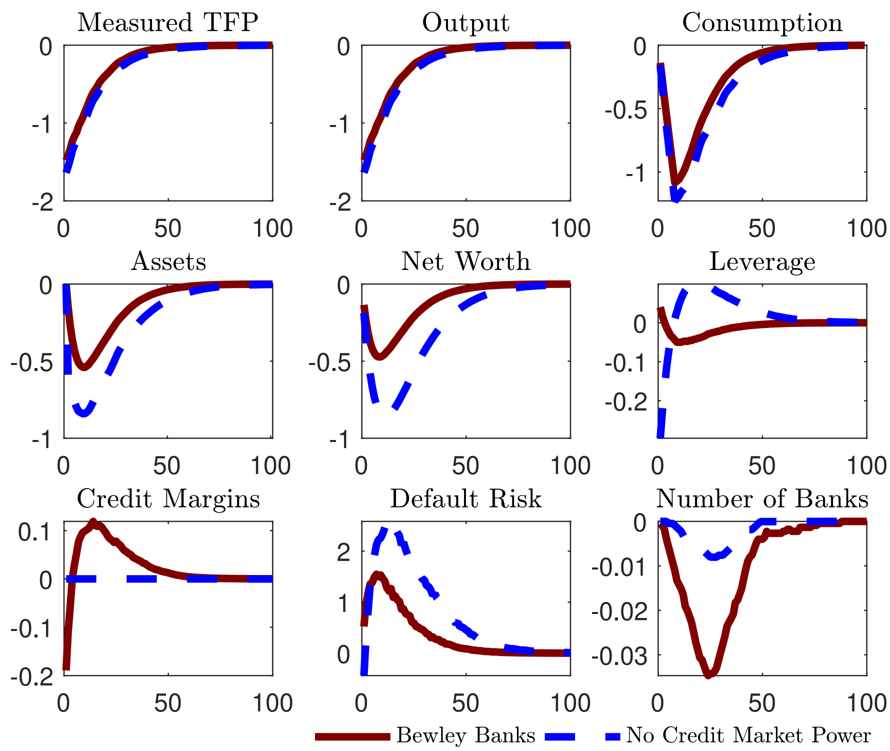
3.5.1 Shutting Down Bank Market Power

In order to analyze the impact of bank market power on business cycle fluctuations, we shut down the credit margins channel. Specifically, we set θ to a very large number. This turns off the bank's ability to charge market-specific loan margins over the cost of funds. However, because the idiosyncratic risk channel and scale-dependency are still active, the distribution of banks still plays a role. Numerically, we run the same exercise as with the baseline crisis experiment but separately for the two economies.

Figure 3.13 plots the response functions to a negative capital quality shock. We observe that the presence of credit market power *dampens* the responsiveness of both output and consumption. In cumulative terms (not shown in the figure), in response to the same shock, the economy with no credit market power suffers a 10% greater decline in output by the 40th period. This result arises from the fact that variable loan margins are an additional margin of adjustment for banks to “insure” against adverse shocks *ex post*. In response to negative aggregate shocks, credit margins rise as the distribution of net worth shifts leftward. The aggregate price of capital *increases*, which acts as an endogenous stabilization mechanism. The market value of bank assets falls by less, and the real economy goes through a more benign contraction.

Note that in the economy with no credit market power, bank book leverage falls on impact albeit recovering and growing over time. Leverage falls initially because book assets fall by more than net worth when credit market power is low. Default risk is also considerably higher in the perfect competition economy. This result resembles the often discussed competition-stability trade-off, but applied to the case of business cycle analysis. Finally, the number of banks declines by more in the baseline Bewley economy. This observation comes as a result of the impact of market power on franchise values. The average bank franchise value is lower when the margins are high. This implies that the participation threshold is higher in the baseline economy with monopolistic banks than in the one with perfectly competitive banks, especially in crisis episodes.

Figure 3.13: **The Role of Credit Market Power**



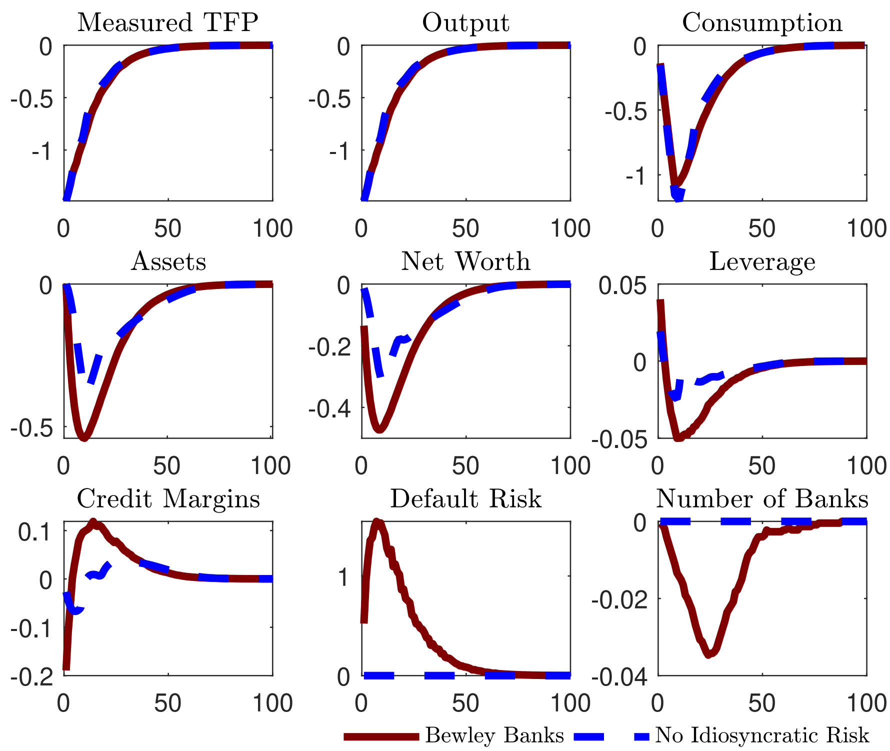
Notes: Response to a one standard deviation negative ψ_t shock. The "Bewley Banks" economy is the baseline. Dashed blue lines represent the economy where θ is set to 10^{10}. All figures plot percentage point differences.

3.5.2 Shutting Down Idiosyncratic Risk

Next we proceed by shutting down the idiosyncratic risk channel. Specifically, we set σ_ξ equal to 0. This has two immediate effects in our economy. First, the economy essentially reduces to the case of a representative financial variety as there is no ex-post heterogeneity in returns. Second, the extensive margin is shut down as well, as a result. Note that in this no-risk economy the bank can still charge credit margins that are market-specific. The details of the numerical experiment are the same as before.

Figure 3.14 plots the responses for the two economies, with and without idiosyncratic risk. Idiosyncratic risk acts as a source of *amplification* in the financial sector with limited effects on output and consumption. This is in contrast to the strong dampening impact of credit market power. The cumulative effect on consumption is around 5-10% greater (in absolute and cumulative terms) in the Bewley economy relative to the no-risk counterfactual. In the Bewley economy, when markets are incomplete, bank assets and net worth decline by more in reaction to aggregate shocks. This effect arises due to ex-post heterogeneity in returns and net worth that the idiosyncratic risk

Figure 3.14: **The Role of Idiosyncratic Risk**

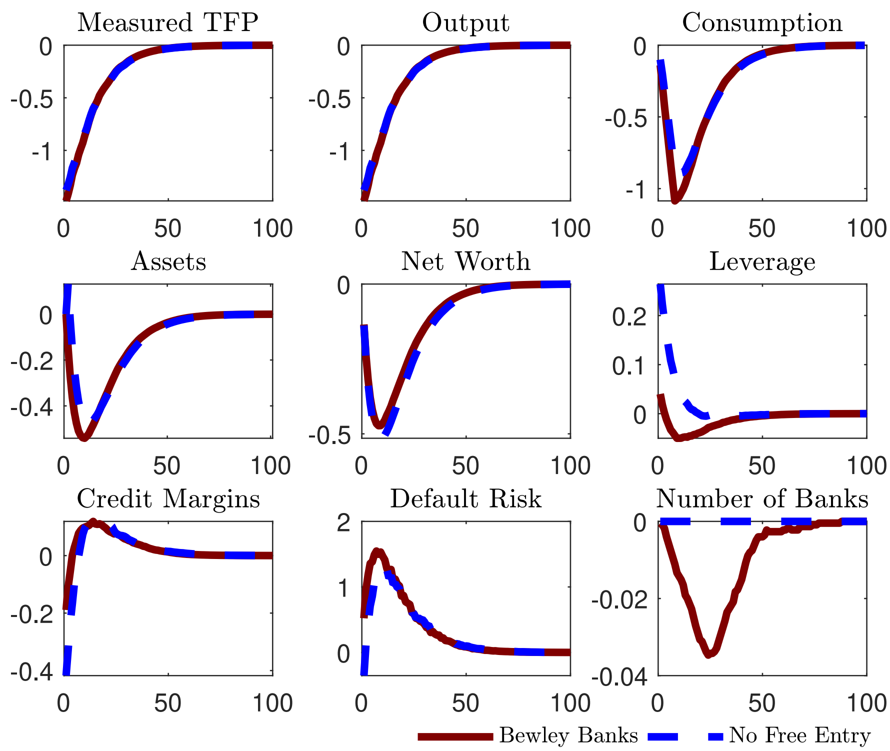


Notes: Response to a one standard deviation negative ψ_t shock. The "Bewley Banks" economy is our baseline. Dashed blue lines represent the economy where σ_ξ is set to 0. All figures plot percentage point differences.

and decreasing returns channels grant us: the distribution of banks in the stochastic steady state features a positive mass of fragile intermediaries - those with low initial net worth and a history of low $\xi(j)$. In crises episodes, fragile intermediaries experience heavier balance sheet losses, which is the intensive margin effect. In addition, the fraction of those fragile intermediaries goes up, i.e. the extensive margin adjustment reinforces the original shock. As a result, both aggregate net worth and capital decline by far more and take longer to replenish.

Note that in the no-risk economy, book leverage falls on impact albeit by a small amount. This is due to the fact that book leverage and net worth are positively correlated and there is no bank net worth heterogeneity in the no-risk case. By similar logic, credit margins fall in the no-risk economy but rise in the Bewley economy. As the distribution of bank net worth shifts to the left, and because private credit margins decline in size, the average margin increases. In the no-risk economy, the margin of the representative banker declines due to the first-order effect of the aggregate shock. Finally, there is no default risk in the no-risk economy by construction.

Figure 3.15: **The Role of Endogenous Entry**



Notes: Response to a one standard deviation negative ψ_t shock. The "Bewley Banks" economy is our baseline. Dashed blue lines represent the economy with exogenous bank entry.

3.5.3 Shutting Down Endogenous Bank Entry

Finally we explore the implications of shutting down the bank entry channel. In the baseline economy, recall that financiers must solve the dynamic problem in (3.24) when deciding whether to enter and to operate. We replace this condition with the assumption that the number of banks remains invariant over the business cycle. In other words, the exact mass of varieties that exits due to the σ shock or endogenous default gets replaced regardless of the state of the business cycle such that J_t is time-invariant. In the Bewley economy, the quantity of capital that new entrants intermediate upon entry is endogenous: it depends on the cutoff level of the idiosyncratic return process, which determines the entry decision of the marginal financier, as well as the startup equity injection. In the no-entry economy, we assume that the mass of entrants intermediates 30% of the aggregate capital stock, which is roughly the ratio that arises endogenously in the Bewley economy. Because entry is now exogenous, the cost of entry becomes a redundant parameter. As a result, this exercise isolates the contribution of the mass of new entrants, and not the capital that they intermediate. Details of the exact numerical experiment are the same as before.

Figure 3.15 plots the responses for the two economies. Endogenous bank entry acts as a minor source of *amplification*: aggregate consumption falls by about 10-15% more (in cumulative terms) in the Bewley economy than in the no-entry version. With exogenous entry, output and the measured TFP fall by slightly less. In the financial sector, exogenous entry is associated with a dampened effect on bank assets, leverage, and margins, primarily due to changes on impact. In the no-entry economy, the volume of capital contributed by new entrants is time-invariant, while it fluctuates together with the mass of new entrants in the Bewley economy. Following an adverse aggregate shock, there is immediately less entry and, by extension, lower capital in the Bewley economy. Over time, the effect vanishes slowly because the banking sector manages to re-grow back the stock of net worth.

The differential effect on capital explains the noticeable differences in the responses of leverage, credit margins, and default risk. With exogenous entry, leverage spikes up on impact because assets initially increase slightly and then fall by less than bank net worth. Credit margins first fall considerably and then rise, with a cumulative growth of about 25% less than in the baseline. This takes place because private margins $\frac{p_t(j)}{\bar{R}(j)}$ are decreasing in aggregate demand for capital K_t. Collectively, this pushes up the average relative price P_t when capital falls by more. Similar logic applies to the response of default risk: $\nu_t(j)$ increases with ψ_t and declines with K_t. As initial aggregate capital changes by less in the no-entry economy, so does default risk.

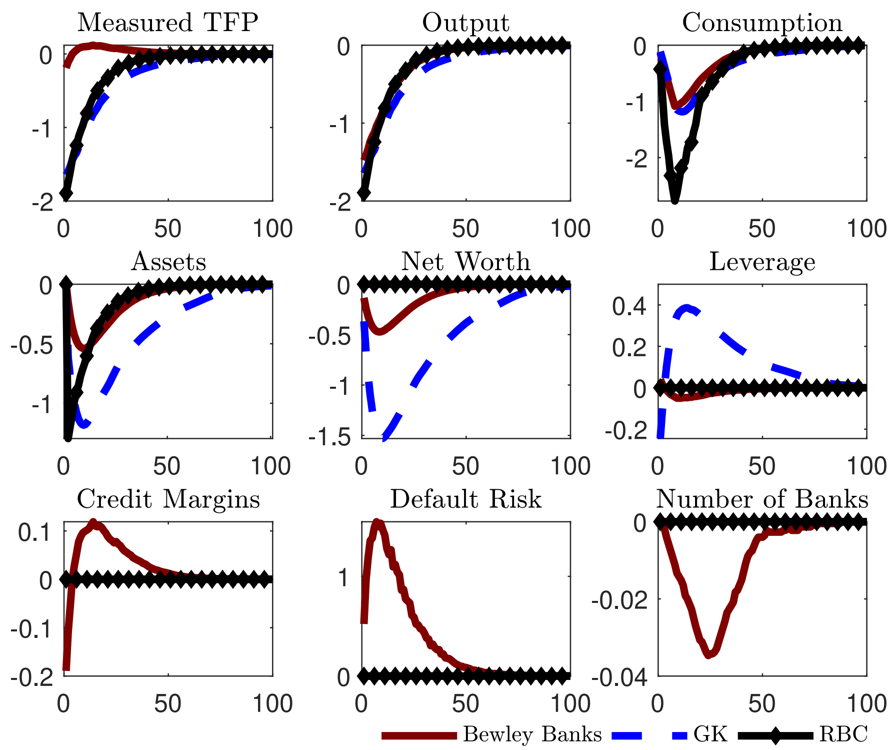
3.6 Main Results and Experiments

In this section we report our main quantitative results. First, we show how our model tractably nests the canonical GK and RBC environments. Recessions in our environment can be either dampened or amplified, depending on the cyclicality of idiosyncratic risk. Second, we demonstrate a specific feature of our environment, i.e., the aggregate state-dependency on the endogenous, dynamic distribution of bank net worth. Third, we use our model to simulate a persistent rise in banking concentration that has been documented for the US. We conclude by identifying and characterizing banking crises episodes in our model economy using an event study approach.

3.6.1 Nesting GK and RBC Models

We now show how our framework nests the RBC and GK models. The GK environment can be achieved in several simple steps. First, we eliminate credit market power by setting θ to a very large number. The distribution of relative bank-level prices p(j) collapses to unity. Second, we set $\beta = 1$ which brings back scale invariance. Finally, we set $\sigma_\xi = 0$ which shuts down the idiosyncratic risk channel. The resulting financial intermediary sector collapses one-to-one to the

Figure 3.16: **Financial Recession - Bewley Banks, GK Banks, and RBC**



Notes: Response to a one-standard deviation negative shock to ψ_t. The red straight, blue dashed, and black diamond lines represent, respectively, the baseline model, the GK counterfactual with no idiosyncratic risk and no monopolistic competition in banking, and the RBC counterfactual with no leverage constraints in addition to all the assumptions from GK. All figures plot percentage point differences.

representative bank in GK. To go from GK to RBC, we set the leverage constraint parameter λ to a very low number. That is, financial frictions are absent and the leverage constraint is always slack at all points of the idiosyncratic and aggregate state spaces.[10] We then compare the response to a one-standard-deviation aggregate ψ shock in the Bewley Banks framework to the GK and RBC models.

Figure 3.16 presents the results. Several observations are worthy of note. First, the GK economy goes through the most contractionary recession due to the financial accelerator embedded into the banking sector. Second, notice that the change in bank leverage is almost an order of magnitude greater in the GK case. Moreover, qualitatively the change is in the opposite direction - leverage falls on impact and then grows over time in the same way as in the full-competition scenario displayed in Figure 3.13. In contrast, market leverage (not shown) increases in both economies but again by almost an order of magnitude more in GK. Last but not least, macroeconomic responses in the

[10] Technically, our "RBC" economy is the true RBC model up to the augmented stochastic discount factor $\bar{\Lambda}$. Recall that the representative investor in our economy is the financial intermediary and not the household.

Bewley economy are *dampened* relative to GK. This result emerges for the following two reasons. Recall that credit market power acts as a significant dampener of exogenous aggregate shocks, a fact that we established in Section 3.5.1. The dampening nature of market power dominates the idiosyncratic risk channel, which amplifies the no-risk counterfactual.

But even in the limiting case with idiosyncratic risk but no market power there is still dampening relative to GK. This occurs because private capital and net worth are both increasing in σ_ξ due to the precautionary lending motive. The only way for the intermediary to hedge idiosyncratic risk is to lend ("save") more. The net effect on market leverage, a sufficient statistic for probability of default, is negative. As a result, an economy with more exogenous environmental risk features less *endogenous* riskiness due to lower leverage ratios and more equity. The inverse relation between risk and leverage is consistent with the conventional wisdom that has been stressed by a number of other papers (Fostel and Geanakoplos, 2008; Gertler et al., 2012). Lower leverage in the stochastic steady state leaves the economy in a less fragile initial condition.

3.6.2 Counter-cyclical Return Risk and Endogenous Amplification

One way to counteract the power of the precautionary lending motive and to enhance amplification is to introduce *counter-cyclical* rate of return risk. As we have shown in Section 3.2.4, there is empirical support for this channel.[11] We now allow idiosyncratic risk to be state-dependent. Specifically, we assume that $\mu_\xi(\mathbf{S})$ - the unconditional mean of idiosyncratic rate of return risk in Equation 3.10 - is now state-dependent and deterministic: it falls by 1 percentage point in the low aggregate state. As a result, when the aggregate state of nature is low, the bank faces a higher probability of experiencing a bad idiosyncratic return draw. This puts additional downward pressure on bank balance sheets precisely when the marginal value of net worth is the highest, thus increasing the unconditional risk premium.[12]

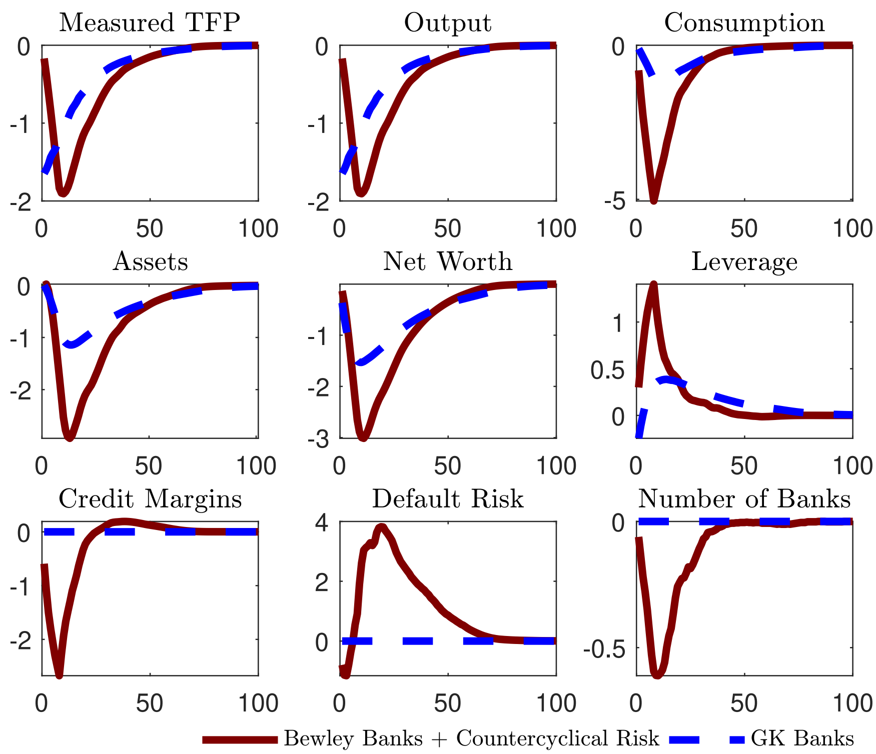
The results of this experiment are portrayed in Figure 3.17. Notice that, on impact, the recession is milder than in the GK benchmark. In fact, it's quantitatively similar to our baseline. However, once in the negative aggregate state, the counter-cyclical idiosyncratic risk channel kicks in. A higher fraction of local credit markets starts to draw low $\xi(\mathrm{j})$. This pushes the distribution of bank assets and net worth leftwards. Over time, the original aggregate shock gets amplified considerably. The measured TFP, output, and consumption all fall by more than 30% relative to GK in cumulative terms by quarter 20.

The amplification effect in the financial sector is considerable. First, the quantitative effect on

[11] Bloom et al. (2018) have shown that microeconomic volatility for non-financial firms rises sharply in recessions. Among many others, Challe and Ragot (2015) and Sterk and Ravn (2020) develop quantitative frameworks with counter-cyclical income and unemployment risk.

[12] Making σ_ξ counter-cyclical would further strengthen the outcome. However, as also noted in Bloom et al. (2018), we would still require a negative shock to the mean in addition to the second-moment disturbance.

Figure 3.17: **Bewley Banks with Counter-cyclical Return Risk σ_ξ**



Notes: Response to a one standard deviation negative ψ_t shock. The "Bewley Banks" economy is our baseline. Dashed blue lines represent the economy where $\mu_\xi(\mathbf{S})$=1 when ψ is high and $\mu_\xi(\mathbf{S})$=0.99 when ψ is low. All figures plot percentage point differences.

the number of active intermediaries is an order of magnitude greater than in the baseline Bewley Banks economy. This result follows immediately from the greater relative fall in both assets and net worth. Second, credit margins now fall both on impact and along the transition back to the original stochastic steady state. This occurs because private margins $\frac{p(j)}{\bar{R}(j)}$ are decreasing in $\mu_\xi(\mathbf{S})$. As a result of the combined effects from book assets and equity, bank leverage also rises on impact considerably.

Overall, we find that in the Bewley Banks framework, recessions could either be dampened or amplified relative to a perfectly competitive benchmark with no idiosyncratic risk. The key determinant is whether the precautionary lending motive can dominate the direct effect of business cycles on financial balance sheets. If the idiosyncratic return risk is counter-cyclical, the business cycle impact on balance sheets is too powerful for the intermediary to insure against ex-ante or ex-post. However, if idiosyncratic risk is not state-dependent, the precautionary lending motive guarantees that recessions are dampened relative to GK as the bank accumulates enough equity capital to withstand aggregate uncertainty.

3.6.3 Fragile Bank Distributions and the Business Cycle

We now highlight a genuine feature of the Bewley Banks environment: namely that aggregate responses to exogenous shocks depend explicitly on the dynamic, endogenous distribution of bank net worth. Our exercise consists of the following steps. First, we solve our model with aggregate uncertainty. For simplicity, the only exogenous aggregate disturbance is the capital quality shock ψ_t. Second, we simulate the model for T=10,000 periods four times. The first simulation is our "benchmark" case - the stochastic steady state. In the second simulation, the model economy is hit with a counterfactually low ψ_t after T* periods. The negative shock is allowed to revert back to the benchmark series with its normal autocorrelation coefficient while being subjected to normal shocks as in the benchmark. The difference between the second and the first simulation is our response function to the negative ψ_t shock. This is also our baseline recession experiment in Section 3.4.5.

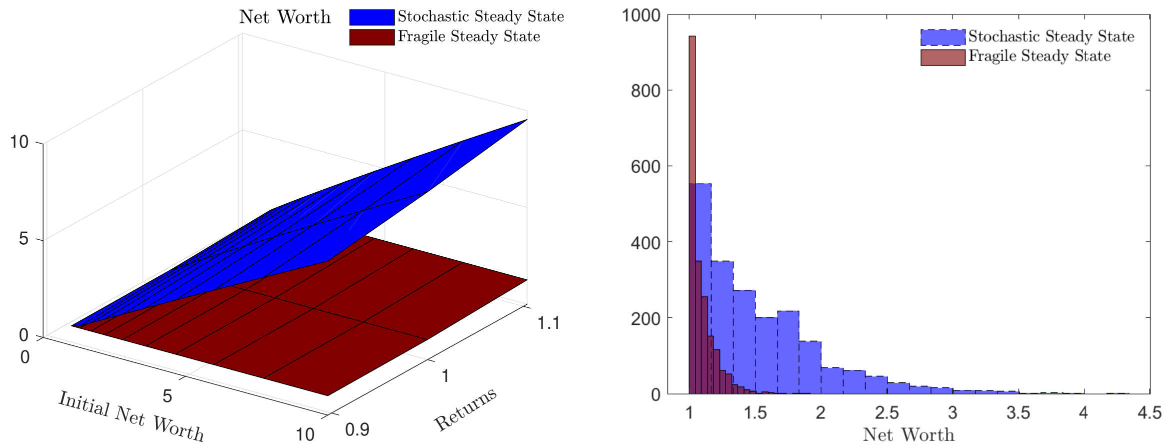
In the third simulation, after T*-1 periods we allow for a transitory exogenous change in the conditional distribution of net worth $n'(j)$. We assume that the economy temporarily moves to a "fragile" distributional state in which average net worth falls while dispersion and skewness increase.[13] We construct a new conditional distribution that is determined by a counterfactual policy function $\hat{n}'(\mathbf{s};\mathbf{S})$ equal to future net worth conditional on the *lowest* value of the initial net worth state. This shift is best visualized in the left panel of Figure 3.18. The blue surface is the equilibrium two-dimensional policy function which we reported in Figure 3.7. The red surface is the policy function consistent with the fragile state. The right panel of Figure 3.18 plots the conditional cross-sectional distributions of net worth from the equilibrium and the fragile states. We assume that this transitory shift lasts for 8 quarters. That is, for 8 quarters the model is being simulated with the counterfactual policy function that generates the fragile distribution. The duration of the shock is consistent with the average duration of banking crises in the data (e.g., Laeven and Valencia (2012)) but is otherwise not materially important. After 8 quarters we revert back to the policy function consistent with the stochastic steady state.

Finally, in the fourth simulation we have the exact same distributional shock as in simulation 3. In addition, in period T* the economy is now hit with the same negative exogenous ψ_t shock as in simulation 2. The difference between the fourth and the third simulations is the model response conditional on the initial banking distribution being fragile. Finally, we compare the percentage differential between simulations 1 and 2 with the percentage differential between simulations 3 and 4. This identifies exactly the aggregate state-dependency of the economy with respect to fluctuations in the bank net worth distribution.

Figure 3.19 reports the results of this exercise. We plot cumulative impulse response functions

[13] In Section 3.6.4 we explore this point from a different angle: by studying model responses to direct, exogenous shocks to higher-order moments which in turn feed into the decline of the first moment endogenously.

Figure 3.18: **The Fragile Steady State**



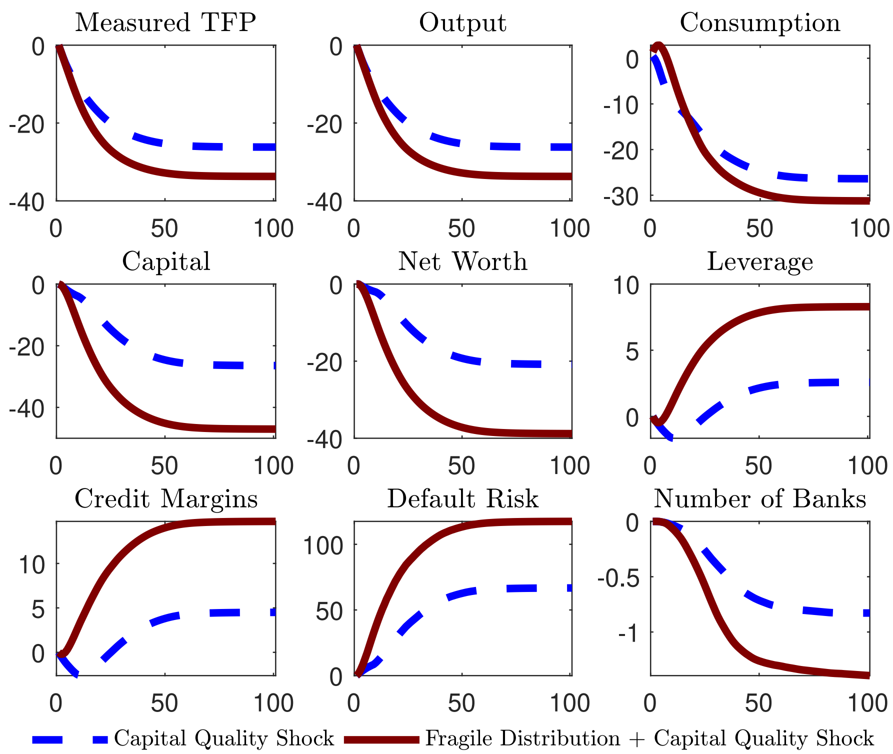
Notes: the left panel shows the baseline and the perturbed two-dimensional policy functions for net worth growth $\hat{n}'(\mathbf{s}; \mathbf{S})$. The right panel plots histograms of the cross-sectional distributions of net worth under the two alternative policy functions. The mean and skewness of the distribution in the stochastic steady state are 1.5151 and 1.4496, respectively. In the fragile steady state, the mean and skewness are 1.0878 and 2.2315, respectively.

for visibility. The dashed blue line is the model response to a negative capital quality shock. The red straight line is the model response to the same shock but conditional on the fragile initial bank distribution. We see from the figure that even a short, transitory negative change in the distribution of bank net worth has a permanent and considerable effect on the macroeconomy. First, the cumulative response of aggregate production is lower by 15% in the case of the fragile distribution. The response of aggregate consumption is lower by roughly 10%. Second, the distributional change has a relative contractionary effect on the size of the financial sector. The negative response of bank assets and net worth is greater in absolute terms by a factor of 2.

Third, the financial sector is more *risky* as the market leverage ratio increases by more than a factor of roughly 2.5. The probability of bank default is higher by a similar order of magnitude. Fourth, the fragile distribution contributes to higher loan margins in the recession. Average margins increase by 10 percentage points more than in the baseline economy. Finally, the number of banks is twice as low, as potential entrants refrain from entry while internalizing low startup equity injections. All of these results arise from the fact that the initial level of equity capital is low - since bank net worth is the key state variable in the model, all endogenous responses such as margins and leverage react accordingly.

Overall, the above exercise highlights the powerful amplification mechanism that is behind the dynamic cross-section of bank net worth. When the initial distribution of net worth is fragile, aggregate responsiveness to negative exogenous shocks is considerably stronger.

Figure 3.19: **Aggregate State Dependency on the Distribution of Bank Net Worth**



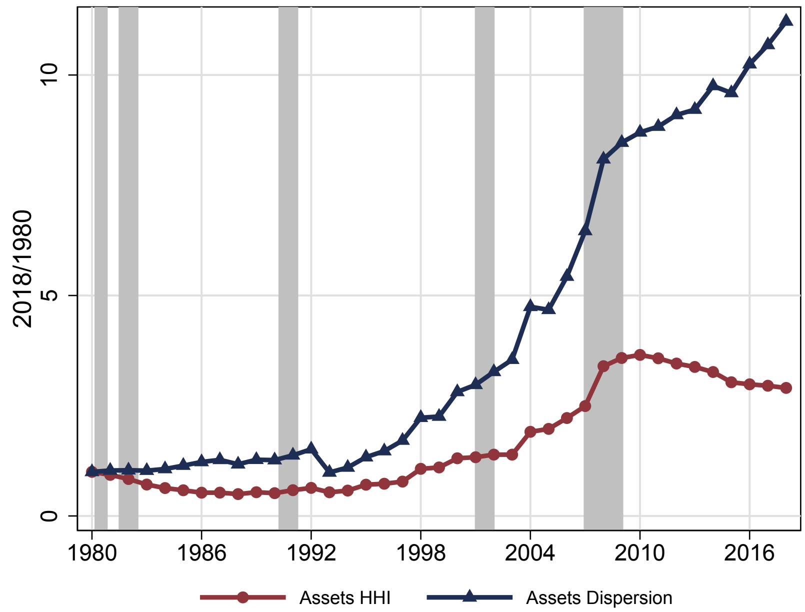
Notes: Impulse responses to a one-standard deviation negative ψ_t shock with and without a prior transitory negative shock to the conditional cross-sectional distribution of bank net worth n(j). The distributional shock lasts for 8 quarters and is depicted in Figure 3.18. The ψ_t shock reverts back to the stochastic steady state with the normal autocorrelation of 0.914. See main text for more details. All figures plot percentage point differences. All response functions are cumulative.

3.6.4 The Rise of Banking Concentration and Dispersion

Recent papers by Jamilov (2020) and Corbae and D'Erasmo (2019) have documented a considerable rise in *concentration* in the U.S. banking sector. There is a large literature in corporate finance and banking that links the rise of banking concentration to a plausibly exogenous sequence of state and federal legislations that relaxed restrictions on bank entry and geographical expansion over 1960s-1980s. (Jayaratne and Strahan, 1996; Kroszner and Strahan, 2014). Figure 3.20 plots the time-series of commercial bank assets concentration (measured by the HHI) and dispersion. The two measures are highly correlated and have risen by a factor of 4 and 10, respectively, since the 1980s.

In our quantitative exercise below, we explore exogenous, transitory but persistent, shocks to the higher-order moments of the bank credit distribution as a *source* of business cycle fluctuations. We operationalise this idea in the following way. Recall that according to our numerical algorithm,

Figure 3.20: **The Rise of Banking Concentration and Dispersion**



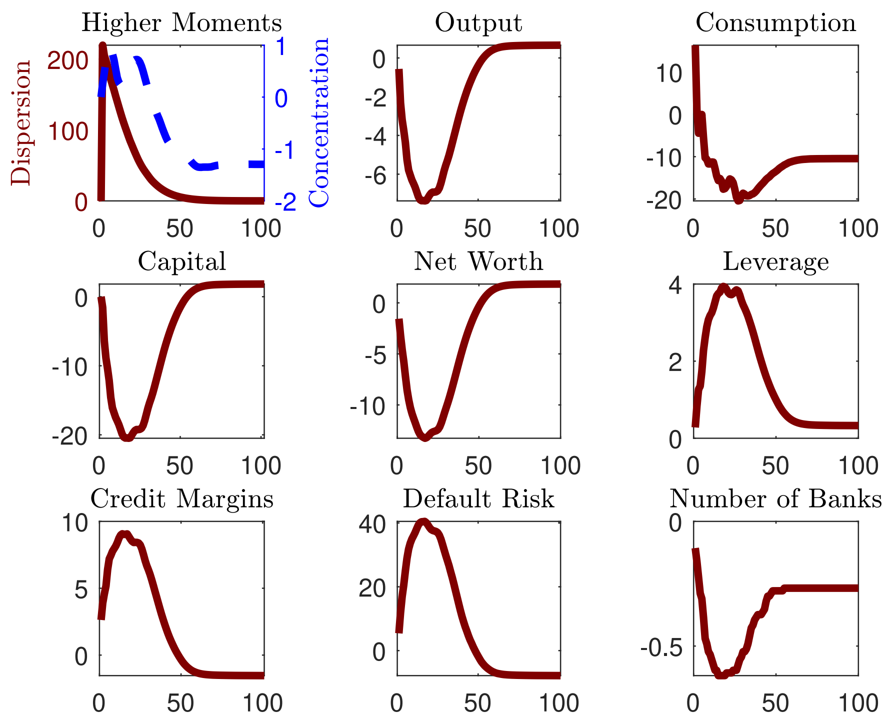
Notes: Time-varying dispersion and Herfindahl index of the cross-section of financial intermediary assets. Data is from Compustat. Sample includes US commercial banks only.

we track m_k moments of the k(j) distribution. In the baseline scenario, we only keep track of the mean. In this section, we also track the time-varying dispersion of assets, σ_t^k. In our numerical experiment, as we show below, concentration will rise endogenously on impact in response to the exogenous dispersion shock.

Computationally, tracking the second moment makes σ_t^k a relevant *state variable*. Exogenous shocks to σ_t^k will have a direct first-degree effect on the optimal responses of all agents. If accuracy of the baseline solution could be improved with the introduction of the second moment, then the mean of k(j) is not a sufficient statistic for the characterization of the dynamic cross-section. This line of reasoning is conceptually very similar to the original ideas in Krusell and Smith (1998) who found that the first moment of the distribution of household wealth was generally sufficient for the description of macroeconomic aggregates.

Our numerical exercise consists of three general steps. First, we solve the model where we allow the law of motion of the distribution Γ to have σ_t^k as an additional argument. Second, we simulate the model for 10,000 periods twice. In the first simulation, the economy remains in the stochastic steady state. In the second simulation, the economy is hit with a positive, persistent shock to σ_t^k in period T*. We assume that σ_t^k rises by a factor of 8, which is in line with empirical evidence in Figure 3.21. Finally, a cumulative impulse response graph plots the percentage differential of the two series. If the second moment is redundant, the response functions should be flat.

Figure 3.21: **Response to an Exogenous Second-Moment Shock**



Notes: Response to an 8-fold positive shock to the time-varying dispersion of the cross-sectional distribution of bank assets k(j). Dispersion reverts back to the stochastic steady state with autocorrelation of 0.9. See main text for more details. Responses are cumulative. All figures plot percentage point differences.

The equilibrium of the model in which σ_t^k is a state variable is characterized by the following log-linear solution for Γ:

$$\begin{aligned}\log(K') &= 0.4465 + 0.8160\log(K) - 0.0010\log\Big(\sigma(K)\Big); \quad \psi \text{ low}\\ \log(K') &= 0.3769 + 0.8526\log(K) - 0.0060\log\Big(\sigma(K)\Big); \quad \psi \text{ high}\end{aligned}$$

For projections of K'. For projections of P' we have:

$$\begin{aligned}\log(P') &= 1.3871 - 0.4406\log(K) + 0.0562\log\Big(\sigma(K)\Big); \quad \psi \text{ low}\\ \log(P') &= 1.4764 - 0.4907\log(K) + 0.0748\log\Big(\sigma(K)\Big); \quad \psi \text{ high}\end{aligned}$$

From the solution above we immediately notice that shocks to σ_t^k are contractionary - they cause quantities to fall and credit margins to rise.

Figure 3.21 plots the impulse response functions. We show strong evidence that persistent shocks to higher-order moments of the banking distribution have a large impact on business cycle fluctuations. First, positive dispersion shocks cause considerable economic recessions: severe cumulative declines in aggregate output, consumption, and the measured TFP. Second, the financial sector goes through a financial crisis as bank assets, net worth, and the number of active institutions all go down significantly. Third, the financial industry accumulates more leverage when dispersion is high, i.e., the economy remains riskier for longer. Fourth, loan margins increase by a factor of 10 in cumulative terms. Overall, we can conclude that the rise of U.S. banking dispersion and concentration over 1980-2020 may have contributed to a more sluggish growth with fewer, smaller and riskier financial intermediaries that charge higher loan margins and default more often.

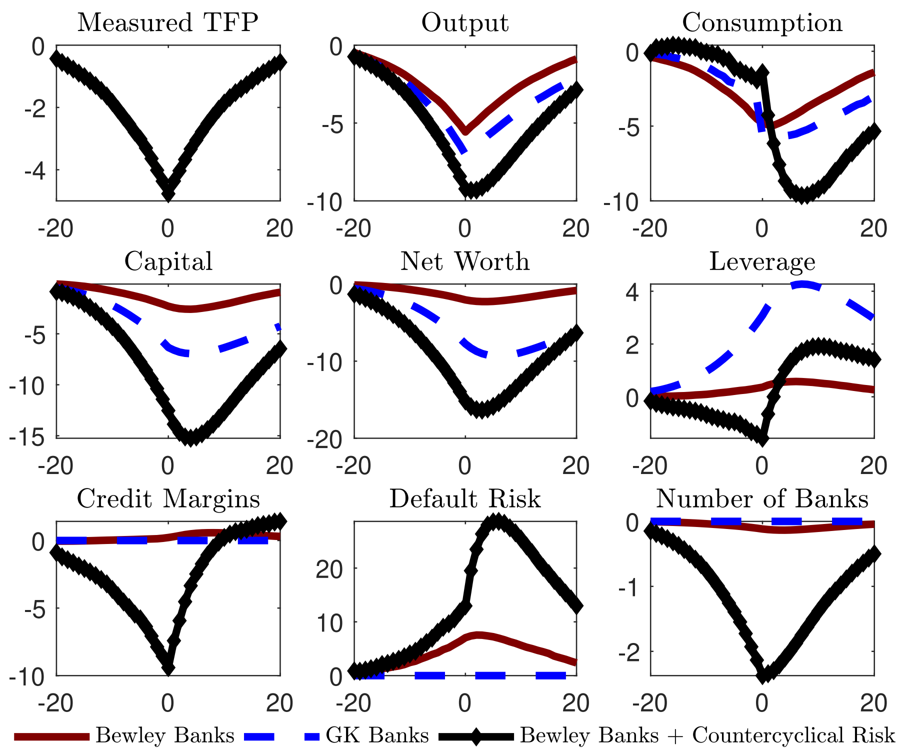
3.6.5 Banking and Economic Crises

In this section we use our model to identify and characterize systemic banking and economic crises. We employ an event study approach. Our methods follow the open-economy macroeconomics literature (Mendoza, 2010). First, we solve the model and simulate it for 10,000 periods with ψ_t as the only exogenous aggregate disturbance. Second, we define economic crises as episodes (quarters) with low measured TFP, $\tilde{A}_t$. Specifically, $\tilde{A}_t$ must be in the lower 20% of the whole simulation. This approach enables a fair comparison across different classes of models because $\tilde{A}_t$ can be readily constructed in both GK and Bewley Banks frameworks. If the bank default risk channel is active, our definition parsimoniously captures episodes of joint financial and economic distress because default risk in our framework is countercyclical.[14] Third, we store every crisis episode and look at the 20-quarter window before and after the event. For each quarter in the window, we calculate the unweighted average of key macroeconomic and financial aggregates. Finally, we perform the same exercise for the GK economy and for the Bewley economy with counter-cyclical idiosyncratic risk.

Figure 3.22 reports the results. In every picture, the red straight line corresponds to the Bewley economy, the blue dashed line corresponds to GK, and black diamond markers to the Bewley economy with counter-cyclical idiosyncratic risk. Note that the decline in $\tilde{A}_t$ across the three economies is equalized by construction. We first observe that the two baseline economies - Bewley and GK - go through the same crisis in very different ways. At the peak of the crisis, the Bewley Banks economy features a smaller contraction in output, consumption, bank assets, bank net worth. The decline in the number of active intermediaries is also relatively muted. In the Bewley economy, crises are also characterized by a slight build-up of loan margins in the pre-event phase. Meanwhile, at the peak of the crisis the Bewley economy displays a higher default risk but a lower bank leverage ratio. This is a variant of the canonical financial competition-stability trade-off. This observation

[14]The literature on banking crises has documented the substantial negative impact of bank defaults on the real economy and consumer welfare, which in our exercise is captured in a general way with low measured TFP (Laeven and Valencia, 2012).

Figure 3.22: **Banking and Economic Crises - Event Study Approach**



Notes: Event study analysis of an economic crisis that is defined as a quarter with measured TFP in the bottom quintile of the 10,000 period-long simulation. Straight red, dashed blue, and black diamond lines represent, respectively, the baseline model, the GK counterfactual with no idiosyncratic risk and with perfect banking competition, and the Bewley economy with counter-cyclical idiosyncratic risk. All figures plot percentage point differences.

is also consistent with our impulse-response experiment in Figure 3.16.

Now, consider the behavior of the Bewley economy with *counter-cyclical idiosyncratic return risk*. Economic crises are associated with considerably more severe contractions in output and consumption. The deterioration in the financial sector activity is far more severe relative to the GK counterfactual. Bank assets and net worth fall by 5-10 percentage points more. Furthermore, bank default risk is significantly amplified. That is, economic recessions in the Bewley economy with counter-cyclical risk are far more likely to occur jointly with systemic banking crises and episodes of financial instability and fragility. Finally, the number of active intermediaries falls by an order of magnitude more than in the baseline Bewley economy.

It is important to highlight that our crisis episodes are still characterized by declines in aggregate risk, in all three economies. That is, our result points to the *amplification* of contractionary shocks, rather than on the endogenous build-up of risk during booms. A key reason for this distinction is the structure of our model that yields *counter-cyclical* market leverage due to the assumed specification of the equity-based leverage constraint.

3.7 Alternative Aggregate Shocks

So far the sole source of aggregate uncertainty in our baseline economy has been a capital quality shock ψ_t. In this section we investigate the extent to which alternative aggregate shocks can match the business cycle statistics reported in Table 3.1. We explore six potential "candidate" shocks. Shock by shock, we re-solve our model under the assumption that it is the only source of aggregate uncertainty in the environment. We then simulate the model economy for 10,000 quarters and report time-series correlations between Y_t and our variables and moments of interest.

The six shocks that we consider are the following. First, the baseline shock to the quality of aggregate capital ψ_t. Following Merton (1973), this shock captures fluctuations in the value of capital - its sudden obsolescence or valuation. Second, we consider a shock to Hicks-neutral total-factor productivity A_t (Kydland and Prescott, 1982). This is a standard exogenous stochastic component in the Real Business Cycle literature. Third, a shock to the banks' dividend payout ratio σ_t. This shock essentially captures changes in consumer "preferences" as was originally proposed in Krusell and Smith (1998). Our approach is more parsimonious because we do not consider shocks to β, as the authors did, for simplicity: σ_t only has a direct effect on the augmented SDF of the intermediary and an indirect effect on household behavior through general equilibrium channels.

The fourth shock we consider is a disturbance in the leverage constraint parameter λ_t. In the literature, this is sometimes labeled as a "financial shock" as in Jermann and Quadrini (2013) and Khan and Thomas (2013). The shock captures sudden changes in the ability of banks to borrow and build up leverage. The fifth shock we consider is the "credit markup shock" θ_t in the spirit of Clarida et al. (2002) or Ball et al. (2005). This shock can capture sudden changes in the degree of financial market competition and concentration. Finally, a shock to the degree of insurability of idiosyncratic shocks κ_t. This shock best resembles the "incompleteness shock" of Davila and Philippon (2019) and captures sudden disruptions in financial market trading, for example a liquidity dry-out.

Table 3.5 presents correlations with output of the first three moments of model-implied time-varying distributions of bank assets $k_t(j)$, net worth $n_t(j)$, leverage $\phi_t(j)$, loan margins $\chi_t(j)$, and insolvency risk $\nu_t(j)$. We also include the mass of entering varieties M_t and the total number of banks J_t.

Consider the performance of the baseline (capital quality) ψ_t shock. As discussed before, this shock nails down all but two moments: dispersion of margins and concentration of leverage. Shocks to TFP do rather poorly. Specifically, they predict counter-cyclical book leverage mean and dispersion. They also miss the cyclicality of credit margins. Shocks to the elasticity of substitution θ_t ("credit markup shocks") and the dividend payout σ_t generate similar responses: both cannot match the counter-cyclicality of the concentration of assets. In addition, shocks to θ_t incorrectly

Table 3.5: **Business Cycle Statistics - Alternative Shocks**

Correlation with GDP of	Data	Capital Quality (ψ_t)	TFP (A_t)	Leverage Constraint (λ_t)	Credit Margins (θ_t)	Dividend Payout (σ_t)	Market Incompleteness (κ_t)
Assets Mean	0.498	0.798	1.000	1.000	1.000	1.000	1.000
Assets Dispersion	0.642	0.541	0.974	0.944	0.604	0.661	0.476
Assets Concentration	-0.568	-0.103	-0.040	-0.191	0.104	0.132	0.101
Net Worth Mean	0.211	0.842	0.958	0.547	0.943	0.727	0.931
Net Worth Dispersion	0.544	0.683	0.907	0.536	0.701	0.620	0.449
Net Worth Concentration	-0.472	-0.238	-0.874	-0.470	0.028	0.011	0.009
Margins Mean	-0.563	-0.305	0.035	-0.291	-0.813	-0.809	-0.347
Margins Dispersion	-0.370	0.437	0.565	-0.100	-0.019	0.185	0.010
Margins Concentration	0.725	0.476	0.735	0.614	0.023	0.002	0.065
Default Mean	-0.325	-0.740	-0.592	-0.029	-0.060	0.165	-0.036
Default Dispersion	-0.309	-0.493	-0.031	-0.003	-0.059	0.133	-0.036
Default Concentration	0.033	0.278	0.336	0.377	0.012	-0.039	0.027
Book Leverage Mean	0.701	0.197	-0.357	0.277	0.804	0.621	0.612
Book Leverage Dispersion	0.043	0.097	-0.347	0.142	0.223	0.424	0.209
Book Leverage Concentration	-0.641	0.043	0.639	0.340	0.095	0.219	0.121
Bank Entry Mass	0.700	0.810	0.717	0.119	0.841	0.665	0.858
Number of Banks		0.811	0.726	0.108	0.839	0.669	0.840

Notes: Data- and model-implied correlations with output Y_t. Each column reports correlations based on an economy with a single source of aggregate uncertainty reported in row 1. In each case, the model is solved and simulated for 10,000 quarters. Correlation between bank entry and output is taken from Corbae and D'Erasmo (2019). See main text for more details.

predict that the concentration of margins is virtually acyclical. Market incompleteness shocks κ_t fail to generate counter-cyclical concentration of assets and net worth as well as counter-cyclical dispersion of loan margins. The most interesting candidate is the financial shock λ_t. It can match all but one moment - counter-cyclicality of the skewness of leverage, which is in fact the only moment that none of the candidates can replicate. A positive shock to λ_t generates a recession by tightening the limit on bank leverage, which in turn reduces the supply of credit to the economy. λ_t shocks are designed to somehow represent fluctuations in the health of the financial sector. They capture exogenous variations in the degree of moral hazard between the lender and the borrower. Overall the dynamic cross-section of financial intermediaries can be best approximated by shocks to the financial system, proxied in our case by shocks either to banks' capital quality or their leverage constraints.

3.8 Conclusion

We have developed a new tractable, dynamic stochastic general equilibrium framework with monopolistic competition and uninsurable idiosyncratic return risk in the financial sector. Our setup builds on the canonical macro-banking models of Gertler and Kiyotaki (2010) and Gertler and Karadi (2011) and nests them as special cases. The simultaneous assumptions of local decreasing returns to scale and idiosyncratic return risk break scale invariance. Because the marginal value of net worth and optimal leverage ratios are now both size-dependent, a time-varying distribution of bank characteristics emerges. With aggregate uncertainty, the distribution of bank net worth becomes a *time-varying* endogenous state variable.

Our framework rests upon four quantitative forces. First, the cyclicality of the idiosyncratic risk process determines whether aggregate contractions in the model economy are dampened or amplified. In the baseline scenario with acyclical risk, the precautionary lending motive dominates the first-degree impact from business cycles and dampens the effects of negative aggregate shocks. However, if idiosyncratic risk is counter-cyclical, the direct effect of business cycle fluctuations on bank balance sheets dominates. As a result, economic recessions get substantially amplified. Second, the model features an explicit aggregate state dependency on the dynamic cross-section of bank assets. Varying the initial state of the distribution - either by targeting the conditional distribution of net worth or through direct exogenous shocks to higher-order moments - has implications on business cycle fluctuations and the aggregate sensitivity to exogenous shocks.

Third, individual banks do not internalize the impact of private loan margin-setting choices on aggregate demand. This is the canonical aggregate demand externality of Blanchard and Kiyotaki (1987) applied to the case of financial intermediaries. Finally, the model generates endogenously the financial competition-stability trade-off. In the Bewley banks framework with acyclical idiosyncratic risk, it is generally the case that severe economic recessions are accompanied by relatively mild financial crises. When idiosyncratic risk is counter-cyclical, economic recessions occur jointly with significant deterioration in the financial sector and elevated financial fragility levels.

Our Bewley Banks framework is tractable and portable. It is possible to introduce nominal rigidities into our model, study unconventional credit policies such as bank-level equity injections, or to relax the closed economy assumption.[15] The tractability of our approach rests on the long and vast literature on monopolistic competition with CES aggregators a la Dixit and Stiglitz (1977). Extensions of the model to generate heterogeneous, variable equilibrium markups would achieve a three-dimensional idiosyncratic state which would include bank net worth, returns, and market

[15]In Jamilov and Monacelli (2020) we study the monetary policy transmission mechanism in a Bewley Banks environment with nominal rigidites.

power. This feature would yield a positive correlation between bank size and loan margins, potentially an interesting and powerful additional channel of transmission. We leave all these interesting and important avenues for future research to explore.

References

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

——— (2011): "Financial Intermediaries and Monetary Economics," *Handbook of Monetary Economics, ed. by Ben Friedman and Michael Woodford*, 3A(12).

AGARWAL, S., R. CORREA, B. MORAIS, J. ROLDAN, AND C. RUIZ-ORTEGA (2020): "Owe a Bank Millions, the Bank Has a Problem: Credit Concentration in Bad Times," *International Finance Discussion Papers*, 1288.

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

ALLEN, F. AND D. GALE (1998): "Optimal Financial Crises," *Journal of Finance*, 53(4).

——— (2004): "Financial Intermediaries and Markets," *Econometrica*, 72(4).

ANOLLI, M., E. BECCALLI, AND G. BORELLO (2015): "Are European banks too big? evidence on economies of scale," *Journal of Banking and Finance*, 58.

BALL, L., G. MANKIW, AND R. REIS (2005): "Monetary policy for inattentive economies," *Journal of Monetary Economics*, 52.

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

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

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

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

BLOOM, N., M. FLOETOTTO, N. JAIMOVICH, I. SAPORTA-EKSTEN, AND S. J. TERRY (2018): "Really Uncertain Business Cycles," *Econometrica*, 86(3).

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

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

CARROLL, C. D. (2006): "The method of endogenous gridpoints for solving dynamic stochastic optimization problems," *Economics Letters*, 91.

CHALLE, D. AND X. RAGOT (2015): "Precautionary Saving Over the Business Cycle," *Economic Journal*, 126.

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

CLARIDA, R., J. GALI, AND M. GERTLER (2002): "A simple framework for international monetary policy analysis," *Journal of Monetary Economics*, 49.

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

CORBAE, D. AND P. D'ERASMO (2019): "Capital Requirements in a Quantitative Model of Banking Industry Dynamics," *NBER Working Paper*, 25424.

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

CUCINIELLO, V. AND F. SIGNORETTI (2015): "Large Banks, Loan Rate Markup, and Monetary Policy," *International Journal of Central Banking*, 11(3).

DAVILA, E. AND T. PHILIPPON (2019): "Incompleteness Shocks," *Working Paper*.

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

DE LOECKER, J., J. EECKHOUT, AND G. UNGER (2020): "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, Forthcoming.

DIEZ, F., D. LEIGH, AND S. TAMBUNLERTCHAI (2018): "Global Market Power and its Macroeconomic Implications," *IMF Working Paper*, 18/137.

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

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

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

GERALI, A., S. NERI, L. SESSA, AND F. SIGNORETTI (2010): "Credit and Banking in a DSGE Model of the Euro Area," *Journal of Money, Credit, and Banking*, 42(s1).

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.

GERTLER, M., N. KIYOTAKI, AND A. QUERALTO (2012): “Financial Crises, Bank Risk Exposure and Government Financial Policy,” *Journal of Monetary Economics*, 59, S17–S34.

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

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.

GUVENEN, F., S. OZKAN, AND J. SONG (2014): “The Nature of Countercyclical Income Risk,” *Journal of Political Economy*, 122(3), 621–660.

HELLMAN, T., K. MURDOCK, AND J. STIGLITZ (2000): “Liberalization, Moral Hazard in Banking, and Prudential Regulation: Are Capital Requirements Enough?” *American Economic Review*, 90(1).

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

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

JAMILOV, R. (2020): “A Macroeconomic Model with Heterogeneous Banks,” *Working Paper*.

JAMILOV, R. AND T. MONACELLI (2020): “Monetary Policy with (H)eterogeneous (BANK)s,” *Working Paper*.

JAYARATNE, J. AND P. STRAHAN (1996): “The finance-growth nexus: evidence from bank branch deregulation,” *Quarterly Journal of Economics*, 111.

JERMANN, U. AND V. QUADRINI (2013): “Macroeconomic Effects of Financial Shocks,” *American Economic Review*, 102(1), 238–271.

KHAN, A. AND J. THOMAS (2013): “Credit Shocks and Aggregate Fluctuations in an Economy with Production Heterogeneity,” *Journal of Political Economy*, 121(6), 1055–1107.

KORINEK, A. AND M. NOWAK (2016): “Risk-Taking Dynamics and Financial Stability,” *Manuscript*.

KROSZNER, R. AND P. STRAHAN (2014): “Regulation and Deregulation of the U.S. Banking Industry: Causes, Consequences, and Implications for the Future,” *N. Rose (eds.), Economic Regulation and Its Reform: What Have We Learned?*, NBER Books.

KRUSELL, P. AND A. SMITH (1996): “Income and Wealth Heterogeneity, Portfolio Choice, and Equilibrium Asset Returns,” *Macroeconomic Dynamics*, 1, 387–422.

——— (1998): “Income and Wealth Heterogeneity in the Macroeconomy,” *Journal of Political Economy*, 106, 867–896.

KYDLAND, F. E. AND E. C. PRESCOTT (1982): “Time to Build and Aggregate Fluctuations,” *Econometrica*, 50.

Laeven, L. and F. Valencia (2012): "Systemic Banking Crises Database: An Update," *IMF Working Paper*, 12/163).

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, 42–49.

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

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

Mendoza, E. G. (2010): "Sudden Stops, Financial Crises, and Leverage," *American Economic Review*, 100.

Merton, R. C. (1973): "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 41.

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

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

Paravisini, D., V. Rappoport, and P. Schnabl (2020): "Specialization in Bank Lending: Evidence from Exporting Firms," *Working Paper*.

Rios Rull, V., T. Takamura, and Y. Terajima (2018): "Banking Dynamics and Capital Requirements," *Manuscript*.

Sterk, V. and M. Ravn (2020): "Macroeconomic Fluctuations with HANK and SAM: an Analytical Approach," *Journal of the European Economic Association*, Forthcoming.

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

Appendix

3.9 Empirical Appendix

3.9.1 Data Description

We acquire financial intermediary balance sheet data for the U.S. (total assets and net worth) from Compustat North America – Fundamentals Quarterly. We include all institutions belonging to SIC sectors "Finance, Insurance and Real Estate" (all codes beginning with 6). We use the variable "ATQ – Total Assets Quarterly" for total assets, "CEQQ – Common Equity Quarterly" for net worth and we compute leverage at the institution-quarter level as the ratio of total assets over net worth. For robustness, we also consider an additional measure of net worth, "SEQQ – Stockholders Equity Quarterly" and results are not affected in a material way. Using the data on assets and equity, we construct the book leverage ratio as the ratio of bank assets over equity. Our baseline sample for total assets, net worth and leverage is over 1985Q1-2020Q1. We experiment with alternative sample durations in Section 3.9.4.

In order to construct bank-level measures of the loan margin, we extract data on interest and non-interest revenues and expenses. We use the Compustat Banks – Fundamentals Quarterly database due to better coverage of income-statement data. As a result, our measure of loan margins is computed only for institutions belonging to the 2-digits SIC sector 60 (Depository institutions).

All throughout, our sample includes companies with headquarters located in the US. We exclude companies that report earnings in any other currency except the USD. All variables are deflated using the U.S. GDP implicit price deflator published in the OECD Main Economic Indicators.[16] We clean the sample from observations that are either erroneous or are extremely outliers. Specifically, we drop observations with book leverage above 100 or smaller than 1 as well as all cases of negative net worth (equity). For our main analysis we focus only on institutions appearing in the dataset for at least 80 quarters. We present robustness checks for different sample definitions in Section 3.9.4.

We construct four proxies for loan margins. Our main measure of margins is computed as the ratio of Total Interest and Related Income (IDITQ) over Total Interest and Related Expenses (XINTQ). For robustness, we also compute the ratio between Net Interest Income (NIINTQ) and Total Interest and Related Expenses. Because of data availability, both of these measures are only available for 1993Q1-2020Q1. In addition, we construct the two following proxies: Net Current Operating Earnings (NCOEQ) divided by Total Interest and Related Expenses and Interest and Fees on Loans (IDILBCQ) over Total Interest and Related Expenses. These measures are available for the full sample 1985Q1-2020Q1. All four definitions yield quantitatively very similar business

[16]We download this measure from the St Louis Federal Reserve, series USAGDPDEFQISMEI.

cycle correlations and volatilities. We prefer the first measure because it is the most complete and captures all relevant factors related to either income or expenses.

In order to construct a bank-level measure of bank default risk, we use the Markit database on Credit Default Swaps (CDS). Our baseline measure is the 5-year CDS spread because it is the most liquid among all the maturities. If we consider 6-months, 1-year or 10-year CDS spreads, results do not change. Our sample refers to CDS contracts that are issued and traded in US Dollars. We restrict reference entities to the "Financial" sector in the US. Our sample runs from 2002Q1 until 2020Q1. For each institution in the sample, we construct quarterly aggregates from daily CDS data.

Before executing our main empirical exercises, we truncate the sample at the 1 and 99 percentiles for leverage, margins and default risk. We do not perform any truncation for total assets and net worth. Generally, truncation of balance sheet items does not affect the results. Truncation of leverage, margins, and CDS spreads helps tighten up correlations of higher-order moments which are usually affected by extreme outliers.

Throughout the paper, we work with the first three moments of the time-varying distributions of intermediary assets, net worth, leverage, loan margins, and CDS spreads. Unless stated otherwise in the text, we always proxy the first moment with the unconditional mean and the second moment with the standard deviation. Our proxy for concentration depends on the variable at hand. For assets and equity we calculate the Herfindahl index (HHI) and for leverage, loan margins, and CDS spreads we compute statistical skewness.[17] Unless stated otherwise, we also always log-linearly detrend all those moments before computing any statistic. To compute business cycle correlations, we log-linearly detrend the quarterly real GDP of the U.S. which we obtain directly from the St. Louis Federal Reserve Board.

As for countries other than the U.S., we acquire financial intermediary balance sheet data from Compustat North America – Fundamentals Quarterly for Canada and from Compustat Global – Fundamentals Quarterly for all the others. We use the same procedures followed for U.S. data with two exceptions. First, we use "SEQQ – Stockholders Equity Quarterly" as our main proxy of net worth, since this is the only variable widely available for all institutions. As a result, we also compute leverage based on this variable. Second, to avoid dealing with an unbalanced number of observations across quarters, we drop from our sample all those institutions that never report for two consecutive quarters. That is, as long as a company has reported for two consecutive quarters at least once, we keep this company in our sample.

Because of data availability, our samples span different time periods according to the reference country. In particular, data for Australia cover 1997Q1-2019Q4, for Canada 1991Q1-2020Q1, for

[17] We compute HHI for variable x according to the usual formula: $\mathrm{HHI}_t(x) = \sum_i \left(\frac{x_{it}}{x_t}\right)^2$.

France 2001Q4-2019Q4, for Germany 2005Q4-2019Q4 and for U.K. 1996Q4-2020Q1.

For CDS we follow again the same steps illustrated for U.S. data. We do not report CDS correlations for Australia, since Markit does not cover Australian companies. Our data cover the same time period as for the U.S., that is 2002Q1-2020Q1, with the exception of Canada, for which the sample starts in 2002Q3.

For these countries, we do not construct a measure of margins, because of insufficient coverage of income-statement data.

3.9.2 Heterogeneity by Intermediary Sub-Sector

As mentioned in the main text, there is multi-modality in the distribution of financial intermediaries. It is important to complement our analysis of the aggregate financial sector with industry-level decomposition. We therefore now perform the same statistical exercise but for each of the six major sub-industries of the broader financial sector: depository credit institutions, non-depository credit institutions, broker-dealers, insurance companies, real estate companies, and holdings and investors. As before, we focus on the U.S. only. Due to data limitations, we cannot construct measures of credit margins or default risk for individual industries. Table 3.6 reports the results. Immediately, we notice that there is considerable heterogeneity across sectors. We can summarize all of the notable observations in three general points. First, and perhaps most interestingly, for some sectors like depository institutions and real estate agents the mean of assets and net worth is *counter-cyclical*. Second, although leverage of the aggregate sector is pro-cyclical, it is in fact counter-cyclical for the depository institutions and non-depository credit-granting institutions. This result is consistent with the empirical evidence and mechanisms discussed in He et al. (2016). Third, as is the case with the aggregate sector, concentration of balance sheet characteristics is almost always counter-cyclical.

We now briefly elaborate more on the role of industry heterogeneity and how it relates to our structural model. Our model features a single aggregated financial sector. On the other hand, empirically, it may be more fruitful to look at the data sector-by-sector. In order to facilitate a fair comparison, we will target the aggregate financial sector as our benchmark for data-model correspondence. However, we acknowledge that business cycle fluctuations could differ noticeably across sub-sectors along the within-sector intensive margin and the extensive margin, i.e., fluctuations in the size of each subsector across time.[18]

[18]Our model could be readily extended to include multiple sectors. For example, a fraction of productive capital could be intermediated by intermediaries that face a value-at-risk constraint, i.e., "broker-dealers". The remaining capital would be managed by investors that face an equity-based constraint on risk-taking. In equilibrium, the two sectors would deliver pro-cyclical and counter-cyclical market leverage, respectively. The extensive margin, which could be endogenized, would be crucial in this setup. However, coupled with credit market power and idiosyncratic rate of return risk, this extension is currently computationally infeasible and is beyond the scope of this paper. It is,

Table 3.6: **Business Cycle Correlations - U.S. Data by Sub-Sector**

	Mean of	Dispersion of	Concentration of	Mean of	Dispersion of	Concentration of
	Depository Institutions (SIC 60)			Insurance (SIC 63 and 64)		
Assets - GDP	-0.521	-0.093	-0.513	0.826	0.762	-0.300
Net Worth - GDP	-0.454	0.361	-0.219	0.497	0.689	-0.210
Leverage - GDP	-0.164	-0.186	-0.121	0.593	0.346	-0.159
	Non-Depository Institutions (SIC 61)			Real Estate (SIC 65)		
Assets - GDP	0.816	0.826	-0.522	-0.435	-0.429	-0.554
Net Worth - GDP	0.663	0.686	-0.441	-0.521	-0.572	-0.740
Leverage - GDP	-0.209	-0.048	0.090	0.251	0.133	0.007
	Brokers and Dealers (SIC 62)			Holdings and Investors (SIC 67)		
Assets - GDP	0.896	0.882	-0.685	0.738	-0.010	-0.836
Net Worth - GDP	0.525	0.454	-0.758	0.608	0.320	-0.703
Leverage - GDP	0.364	0.705	0.403	0.424	0.223	0.165

Notes: For every variable except CDS spreads the sample is 1985q1:2020q1. Every variable has been logged (except the skewness of leverage) and linearly detrended. Bank balance sheet data is from Compustat North America. Industry classification follows the first two digits of the Standard Industrial Classification of economic activities (SIC).

3.9.3 Data on Other Countries

Subject to data availability, we also report business cycle correlations of higher-order moments in other countries. We managed to build reasonably long panels for Australia, Canada, France, Germany, and the United Kingdom. Details on sample construction are in the Appendix. To the best of our knowledge, this is the first attempt to establish robust facts of this type for a set of non-US developed economies. Results are reported in Table 3.7. We can summarize these statistics in three broad points. First, it's most clear that CDS spreads have counter-cyclical first and second (except for Canada) moments, and pro-cyclical skewness. Second, leverage is counter-cyclical for all countries except Australia. There is no systematic commonality for the higher-order moments of leverage. Third, the country that appears to be closest to the U.S. in terms of these business cycle patterns is Australia. Overall, there is substantial degree of heterogeneity across countries for

however, a very fruitful topic for future research.

almost every characteristic.[19]

Table 3.7: **Business Cycle Correlations - Aggregate Data for Different Countries**

	Mean of	Dispersion of	Concentration of	Mean of	Dispersion of	Concentration of
		Australia			Germany	
Assets - GDP	0.207	-0.008	-0.614	-0.142	-0.039	0.012
Net Worth - GDP	0.400	0.170	-0.574	-0.044	0.032	-0.087
Leverage - GDP	0.261	0.106	-0.123	-0.189	-0.151	-0.014
CDS Spreads				-0.329	-0.063	0.112
		Canada			United Kingdom	
Assets - GDP	-0.832	-0.655	-0.109	0.042	0.234	-0.369
Net Worth - GDP	-0.772	-0.751	-0.686	-0.161	0.014	-0.779
Leverage - GDP	-0.805	-0.531	0.790	-0.085	0.124	0.528
CDS Spreads	-0.197	0.026	0.063	-0.527	-0.158	0.455
		France				
Assets - GDP	0.170	0.226	0.053			
Net Worth - GDP	0.156	0.167	-0.302			
Leverage - GDP	-0.122	0.067	0.300			
CDS Spreads	-0.355	-0.322	0.282			

Notes: Every variable has been logged linearly detrended. Bank balance sheet data is from Compustat. CDS data is from Markit. The sample for Australia is 1997q1:2019q4, for Canada is 1991q1:2020q1, for France is 2001q4:2019q4, for Germany is is 2005q4:2020q1, and for the UK is 1996q4:2020q1. See the Appendix for variable definitions and further details.

3.9.4 Robustness

For robustness, we report in Tables 3.8, 3.9, 3.10 and 3.11 below the correlations using different sample definitions. Tables 3.8 and 3.9 show correlations both over the whole sample and by sub-industries when we include all institutions appearing at least 50 quarters (instead of 80, as in the main results) and leave the starting date unchanged to 1985Q1.

Similarly, Tables 3.10 and 3.11 report correlations for the sample starting in 1980Q1 and

[19] Similar to our analysis of cross-industry heterogeneity, a multi-country extension of our framework is possible. The mass of differentiated local credit markets could be re-routed to represent a continuum of countries with ex-ante heterogeneity in size or magnitude of local idiosyncratic riskiness. A single financial intermediary would thus charge country-specific margins over country-specific costs of funds. This extension is on the agenda for future research.

including all institutions appearing at least 80 quarters.

Table 3.8: **Correlations with GDP - US Data, Less Balanced Panel**

	Mean of	St Deviation of	Concentration of
Assets - GDP	0.221	0.631	-0.329
Net Worth - GDP	-0.252	0.511	-0.256
Leverage - GDP	0.670	-0.148	-0.741
Margins - GDP	-0.571	-0.345	0.751
Default Risk - GDP	-0.325	-0.309	0.033

Notes: Aggregate business cycle correlations based on the panel of financial intermediaries that contains only institutions with at least 50 (nonconsecutive) quarters of data over the 1985q1-2020q1 sample. Balance sheet data comes from Compustat. Default risk (CDS) data is from Markit.

Table 3.9: **Business Cycle Correlations - U.S. Data by Sub-Sector, Less Balanced Panel**

	Mean of	Dispersion of	Concentration of	Mean of	Dispersion of	Concentration of
	Depository Institutions (SIC 60)			Insurance (SIC 63 and 64)		
Assets - GDP	-0.762	-0.342	-0.299	0.665	0.745	-0.138
Net Worth - GDP	-0.764	0.131	-0.091	0.194	0.659	-0.063
Leverage - GDP	-0.266	-0.117	-0.378	0.055	-0.139	-0.154
	Non-Depository Institutions (SIC 61)			Real Estate (SIC 65)		
Assets - GDP	0.803	0.817	-0.294	-0.446	-0.385	-0.408
Net Worth - GDP	0.635	0.670	-0.223	-0.437	-0.346	-0.379
Leverage - GDP	0.007	0.093	-0.053	0.267	0.157	0.006
	Brokers and Dealers (SIC 62)			Holdings and Investors (SIC 67)		
Assets - GDP	0.923	0.907	-0.683	0.791	0.228	-0.771
Net Worth - GDP	0.644	0.527	-0.726	0.501	0.286	-0.649
Leverage - GDP	-0.034	0.564	0.362	0.512	0.229	-0.052

Notes: Industry-level business cycle correlations based on the panel of financial intermediaries that contains only institutions with at least 50 (nonconsecutive) quarters of data over the 1985q1-2020q1 sample. Balance sheet data comes from Compustat.

Table 3.10: **Correlations with GDP - US Data, starting 1980**

	Mean of	St Deviation of	Concentration of
Assets - GDP	0.313	0.383	-0.634
Net Worth - GDP	-0.139	0.203	-0.590
Leverage - GDP	0.254	0.077	-0.313
margins - GDP	-0.485	-0.311	0.624
Default Risk - GDP	-0.281	-0.269	0.016

Notes: Aggregate business cycle correlations based on the sample of financial intermediaries that starts from 1980q1. Balance sheet data comes from Compustat. Default risk (CDS) data is from Markit.

Table 3.11: **Business Cycle Correlations - U.S. Data by Sub-Sector, starting 1980**

	Mean of	Dispersion of	Concentration of	Mean of	Dispersion of	Concentration of
	Depository Institutions (SIC 60)			Insurance (SIC 63 and 64)		
Assets - GDP	-0.521	-0.270	-0.606	0.261	0.560	-0.445
Net Worth - GDP	-0.442	0.068	-0.475	-0.155	0.462	-0.412
Leverage - GDP	-0.269	-0.114	-0.103	0.438	0.489	0.242
	Non-Depository Institutions (SIC 61)			Real Estate (SIC 65)		
Assets - GDP	0.580	0.663	-0.544	-0.231	-0.154	-0.332
Net Worth - GDP	0.336	0.448	-0.514	-0.494	-0.440	-0.652
Leverage - GDP	-0.005	0.032	-0.166	0.326	0.253	0.281
	Brokers and Dealers (SIC 62)			Holdings and Investors (SIC 67)		
Assets - GDP	0.871	0.871	-0.743	0.444	0.102	-0.736
Net Worth - GDP	-0.008	0.080	-0.777	-0.027	-0.180	-0.772
Leverage - GDP	-0.005	0.032	-0.166	0.570	0.508	0.389

Notes: Industry-level business cycle correlations based on the sample of financial intermediaries that starts from 1980q1. Balance sheet data comes from Compustat.

3.10 Model Appendix: Proofs

Proof of Proposition 4

The bank solves for each j:

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

The first order condition is

$$\Big(1-\upsilon(j)\Big)R^{T}(j)p(j) + \Big(1-\upsilon(j)\Big)R^{T}(j)\frac{\partial p(j)}{\partial k(j)}k(j) - \bar{R}(j)\left(p(j)\beta k(j)^{\beta-1} + k(j)^{\beta}\frac{\partial p(j)}{\partial k(j)}\right) = 0$$

The elasticity of substitution, ignoring the influence of local credit market-level rates on the aggregate index $P(\mathbf{S})$, is

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

The price-setting rule given marginal costs is

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

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

$$MC(j) := \frac{\beta\theta-1}{\theta}p(j)\frac{\bar{R}(j)}{\Big(1-\upsilon(j)\Big)R^{T}(j)}\left[\left(\frac{p(j)}{P(\mathbf{S})}\right)^{-\theta}K(\mathbf{S})\right]^{\beta-1}$$

Proof of Proposition 5

Guess that the solution to the dynamic problem 3.20 is a value function $V(n(j),\xi(j);\mathbf{S}) = \zeta(n(j),\xi(j);\mathbf{S})n(j)$. Define the default risk-adjusted stochastic discount factor $\tilde{\Lambda}(\mathbf{s};\mathbf{S}) = \Big[\Big(1-\upsilon(j)\Big)\Lambda(\mathbf{S})\Big(1-\sigma+\sigma\zeta(n'(j),\xi'(j);\mathbf{S}')\Big)\Big]$. The solution to the program is a system of equations:

$$\mathbb{E}\left[\tilde{\Lambda}(\mathbf{s'};\mathbf{S'})\left(R^{T'}(j) - \bar{R}(j)k(j)^{\beta-1}\right)\right] = \lambda\varphi(n(j),\xi(j);\mathbf{S})$$

$$\varphi(n(j),\xi(j);\mathbf{S})\left[\zeta(n(j),\xi(j);\mathbf{S}) - \lambda\phi(j)\right] = 0$$

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

$$\zeta(n(j),\xi(j);\mathbf{S}) = \varphi(n(j),\xi(j);\mathbf{S})\zeta(n(j),\xi(j);\mathbf{S}) + \mathbb{E}\Big(\tilde{\Lambda}(\mathbf{s'};\mathbf{S'})\Big)\bar{R}(j)k(j)^{\beta-1}$$

Solving for $\zeta(n(j),\xi(j);\mathbf{S})$ yields

$$\zeta(n(j),\xi(j);\mathbf{S}) = \frac{\mathbb{E}\left(\tilde{\Lambda}(\mathbf{s}';\mathbf{S}')\right)k(j)^{\beta-1}\bar{R}(j)}{1-\varphi(n(j),\xi(j);\mathbf{S})}$$

And the Lagrange multiplier on the leverage constraint is

$$\varphi(n(j),\xi(j);\mathbf{S}) = \max\left[1-\frac{\mathbb{E}\left(\tilde{\Lambda}(\mathbf{s}';\mathbf{S}')\right)k(j)^{\beta-1}\bar{R}(j)}{\lambda\phi(j)},0\right]$$

The result follows from (a) the fact that market leverage is $\phi(j) = k(j)^{1-\frac{1}{\theta}}\left(K(\mathbf{S})\right)^{\frac{1}{\theta}}P(\mathbf{S})n(j)^{-1}$ (b) and the previously defined augmented stochastic discount factor $\tilde{\Lambda}(\mathbf{s};\mathbf{S})$. The guess is verified if $\varphi(n(j),\xi(j);\mathbf{S}) < 1$. Size-dependency is guaranteed by $\beta > 1$ so that each bank with different n(j) and $\xi(j)$ chooses a different leverage ratio $\phi(j)$.

3.11 Model Appendix: Accuracy

In this section we discuss the accuracy of our main numerical algorithm. Our convergence tolerance levels for the household and banking problems are $10e^{-8}$ and $10e^{-5}$, respectively. Deposit market clearing is achieved under the tolerance level of $10e^{-5}$ for the deposit rate on each idiosynractic grid point. Finally, tolerance level for the Krusell-Smith recursion is $10e^{-3}$ for both capital and prices. We perform two exercises. First, we report the R-squared from the Γ projections for K' and P'. With a log-linear form, equilibrium of the model with aggregate uncertainty in ψ_t is characterized by the following equations for good and bad times:

$$\begin{aligned}
\log(K') &= 0.3380 + 0.8588\log(K); \quad R^2 = 0.9746; \quad \psi \text{ low} \\
\log(K') &= 0.3668 + 0.8671\log(K); \quad R^2 = 0.9737; \quad \psi \text{ high}
\end{aligned}$$

For projections of K' and

$$\begin{aligned}
\log(P') &= 1.3871 - 0.4788\log(K); \quad R^2 = 0.9909; \quad \psi \text{ low} \\
\log(P') &= 1.5854 - 0.4946\log(K); \quad R^2 = 0.9923; \quad \psi \text{ high}
\end{aligned}$$

For projections of P'.

We also compute the accuracy measure of Den Haan (2010) for the stationary law of motion for aggregate capital and prices. Using the equilibrium law of motion Γ above, without updating this forecast function, we simulate the full-time series of capital and prices using the ψ_t as the source of stochastic exogenous fluctuations. We compare this simulation with the stochastic simulation time-series where aggregate capital and prices are built from the time-varying distribution. The average percentage error is 1.68% for capital and 0.98% for prices. These numbers are in line with existing studies that are similar to our model's degree of non-linearity and complexity (Khan and Thomas, 2008; Nakamura and Steinsson, 2010; Corbae and D'Erasmo, 2019).

References

CORBAE, D. AND P. D'ERASMO (2019): "Capital Requirements in a Quantitative Model of Banking Industry Dynamics," *NBER Working Paper*, 25424.

DEN HAAN, W. (2010): "Assessing the accuracy of the aggregate law of motion in models with heterogeneous agents," *Journal of Economic Dynamics and Control*, 34, 79–99.

HE, Z., B. KELLY, AND A. MANELA (2016): "Intermediary Asset Pricing: New Evidence from Many Asset Classes," *Journal of Financial Economics*, Forthcoming.

KHAN, A. AND J. THOMAS (2008): "Idiosyncratic shocks and the role of nonconvexities in plant and aggregate investment dynamics," *Econometrica*, 76(2), 395–436.

NAKAMURA, E. AND J. STEINSSON (2010): "Monetary Non-neutrality in a Multisector Menu Cost Model," *Quarterly Journal of Economics*, 125(3).