Bewley Banks*

Rustam Jamilov
University of Oxford

Tommaso Monacelli
Bocconi University

April 10, 2023

Abstract

Using bank-level data, we document new facts on the cross-sectional and business cycle properties of the banking distribution. The asset and deposit distributions feature significant heterogeneity and concentration. Estimated credit markups and deposit markdowns are time-varying and counter-cyclical. Idiosyncratic loan return risk is counter-cyclical due to pro-cyclical skewness. We then develop a dynamic general equilibrium model with heterogeneous financial intermediaries, incomplete markets, two-sided bank market power, counter-cyclical income risk, and aggregate uncertainty. The model generates a bank net worth fluctuation problem analogous to the canonical Bewley-Huggett-Aiyagari-Imrohoğlu setup. Both bank heterogeneity and counter-cyclical risk significantly amplify real and financial aggregate fluctuations. Granular shocks to the top quintile of bank assets generate realistic business cycles without any aggregate shocks and can account for almost all of the variation in output due to idiosyncratic risk. Finally, credit market power amplifies aggregate fluctuations while deposit market power dampens them.

Keywords: business cycles, heterogeneous financial intermediaries, bank market power, banking crises

*We thank the Editor and three anonymous referees for many helpful comments and suggestions that have significantly improved the paper. We thank Dean Corbae, Mark Gertler, Julien Matheron (discussant), Morten Ravn, Erwan Quintin, and seminar participants at multiple venues for useful feedback. Marco Bellifemine has provided outstanding research assistance, well beyond the call of duty. Jamilov thanks the AQR Asset Management Institute and the Wheeler Institute for Business and Development for financial support.

Jamilov: rustam.jamilov@all-souls.ox.ac.uk.
Monacelli: tommaso.monacelli@unibocconi.it.

1 Introduction

The 2007-2008 Financial Crisis has led to a considerable reconsideration of how financial intermediaries (banks, for short) affect the macroeconomy (Gertler and Kiyotaki, 2010; Brunnermeier and Sannikov, 2014). Fast-forward fifteen years, the 2023 regional banking crisis is putting banks back on the center stage. Following the collapse of Silicon Valley Bank and the ensuing panic, aggregate bank lending in the United States has fallen by a record $100 billion in the two weeks of March 2023 alone. This has heightened the possibility of a future financial crisis and of an economic recession.
Understanding the role that banks play in business fluctuations is therefore important. Most of the existing studies, however, have so far focused on a *representative* intermediary in environments with perfect competition. In this paper, we develop a tractable, empirically-motivated, quantitative macroeconomic framework where a dynamic distribution of banks with two-sided market power has first-order effects on aggregate fluctuations.

Based on detailed micro data from U.S. depository institutions, we document *four* facts on the banking distribution, both in the cross section and over the business cycle.

Fact 1: *Size heterogeneity and concentration.* The cross-sectional distributions of bank assets and deposits feature high dispersion and, especially, right-skewness. There is a small mass of extremely large intermediaries that co-exist with a large number of smaller banks. Moreover, size concentration has been rising steadily over the past decades. This evidence speaks outright against the assumption of a representative intermediary which is typically assumed in the literature and complements the exceptions that develop heterogeneous-intermediary frameworks (Corbae and D'Erasmo, 2022; Coimbra and Rey, 2023).

Fact 2: *Granularity.* The distributions of bank assets and deposits follow a Power law. If the "Granular Hypothesis" applies (Gabaix, 2011), then idiosyncratic shocks to banks in the right tail of the size distribution can by themselves generate realistic aggregate fluctuations , both on the real and the financial side of the economy.

Fact 3: *Two-sided market power.* There is evidence of time-varying *market power* on both the asset and the liability sides of bank balance sheets. Estimated credit markups (the spread between the credit rate and the marginal cost), are counter-cyclical, pointing to a pass-through to loan rates that is greater than one-to-one. Estimated deposit markdowns (the spread between the deposit rate and the marginal risk-free rate) are also counter-cyclical, suggesting that the pass-through to deposit rates is also greater than one-to-one. Moreover, bank market power is concentrated in the right tail of the size distribution: large banks charge both higher credit markups and lower deposit markdowns.

Fact 4: *Counter-cylical income risk.* The idiosyncratic rate of return risk faced by banks is *counter-cyclical*. This is due to pro-cyclical left *skewness*. In other words, the skewness of

the distribution of returns becomes more negative in recessions, so that conditional on a negative aggregate state the likelihood of a very low idiosyncratic return draw increases. Also, the first moment of the distribution of returns is pro-cyclical, whereas the second moment is a-cyclical, i.e., flat.

Against this empirical background, we develop a tractable, quantitative dynamic general equilibrium framework with aggregate uncertainty and a rich financial intermediation sector whose function is to invest into risky claims on non-financial firms and to source deposits from households. The banking industry features two core properties and deviations from the standard models. First, we introduce both ex-ante and ex-post bank heterogeneity in the rate of return on loans, in the spirit of Benhabib and Bisin (2018); Benhabib et al. (2019). Ex-ante, there is *permanent* return inequality across banks, which we model with a power law density. Second, markets are incomplete and each bank faces *transitory* uninsured idiosyncratic shocks to the rate of return. The permanent-transitory heterogeneity mixture generates a bank net worth fluctuation problem analogous to the canonical Bewley-Huggett-Aiyagari-Imrohoğlu environment (Bewley, 1977; Huggett, 1990; Aiyagari, 1994; Imrohoglu, 1996). Moreover, transitory risk is *counter-cyclical*, i.e., aggregate state-dependent. In recessions, the first and the third moment of the distribution of idiosyncratic return draws both fall. In other words, both the mean and the skewness of transitory returns are pro-cyclical and banks get exposed to greater downside risk to their portfolios in bad aggregate states.

Second, departing from the perfect competition assumption, we allow banks to charge markups to non-financial firms *and* markdowns on household deposits. Households save either in mutual funds or bank deposits but derive utility from the special liquidity services that deposits provide. Internalizing this effect, banks charge markdowns over the risk-free rate. Because marginal liquidity preferences are aggregate state-dependent, markdowns vary over the business cycle. At the same time, loans to non-financial firms are differentiated, which allows banks to charge markups over their marginal costs. The loan aggregator features a "keeping up with the Joneses" form and yields a fully dynamic aggregate demand elasticity and markup. Moreover, since banks are heterogeneous, both markdown and markup choices explicitly depend on bank size, the distribution of which is itself aggregate state-dependent. In equilibrium, a dynamic distribution of two-sided market power emerges.

Our modeling approach eliminates *scale invariance*: all dynamic choices in the financial sector depend on bank-specific characteristics such as the level of net worth. The resulting equilibrium yields a non-trivial, dynamic distribution of bank assets and deposits. The presence of aggregate risk makes this distribution, in principle, an infinitely-dimensional object and a relevant state variable. We resort to numerical methods and the Krusell and Smith (1996, 1998) algorithms to solve our model fully non-linearly. Under perfectly competitive loan and deposit markets, and in the absence of both permanent and transitory

heterogeneity, our Bewley Banks framework nests the canonical Real Business Cycle model and the Gertler and Kiyotaki (2010), Gertler and Karadi (2011), Gertler et al. (2016), Gertler et al. (2020) (GK henceforth) stream of influential macro-banking models as special cases.

We use our framework to show that the dynamic distribution of imperfectly competitive banks has significant implications for aggregate fluctuations. The model is calibrated to match select moments of the U.S. banking sector and business cycle. Most of our quantitative exercises study conditional model-implied responses to an aggregate Total Factor Productivity (TFP) shock - the only source of aggregate uncertainty in the model. We reach six main results.

First, the model generates realistic, right-skewed ergodic distributions of bank size (assets, deposits, and net worth). Both asset and deposit markets are considerably concentrated, mostly due to the presence of permanent bank returns inequality. In the cross section, there is a positive relationship between bank size, credit markups, and deposit markdowns. The calibrated model also generates financial and business cycle statistics that approximate the cyclical properties of the different moments of the U.S. economy rather well.

Second, bank heterogeneity amplifies business cycles. In order to understand this result, we introduce the notion of Marginal Propensity to Lend (MPL), defined as the bank-level response of lending to a marginal change in net worth. In the baseline economy, MPL is heterogeneous and declining in bank size: smaller banks have a greater elasticity of lending with respect to shocks to net worth. Moreover, the average MPL of our economy is greater than the MPL in a representative-bank benchmark. Heterogeneity introduces a mass of banks that are very large and with a low MPL. However, their share is not sufficiently high enough to counteract the larger mass of small, high-MPL banks. As a result, the average MPL is larger relative to the representative-bank benchmark, and the economy - total bank lending in particular - is more responsive to aggregate shocks.

We further elucidate this point by showing that our model features distributional state-dependency: the transmission of aggregate shocks depends on the degree of ex-ante *financial fragility*. Suppose that a negative aggregate TFP shock hits the economy conditional on a financial shock that shifted, in the previous quarter, the distribution of bank capital ratios (defined as the ratio of net worth to assets) leftward. We find that a negative aggregate shock that occurs once the banking sector is already fragile generates a more severe financial and real-economy contraction. Excess contraction scales with the duration and severity of the prior financial shock. The mechanism for this outcome relies on the MPL heterogeneity logic: the fragile economy features a higher starting average MPL because a greater number of banks are close to zero net worth. As a result, any subsequent negative aggregate shock becomes more detrimental.

Third, *counter-cyclical* idiosyncratic bank return risk amplifies aggregate fluctuations. The effect is quantitatively significant. The response of output, household consumption,

bank assets, and bank net worth to a negative TFP shock is up to *fifty per cent larger* in the economy with counter-cyclical risk relative to the a-cyclical risk baseline case. The intuition for this result is that entering a recession triggers the switch towards a more left-skewed density of idiosyncratic draws: banks are more heavily exposed to downside risk. Once this downside risk materializes, a fraction of banks experience extremely bad portfolio outcomes. Bank lending contracts, production stalls, and consumption falls.

Fourth, the model generates both counter-cyclical credit markups and counter-cyclical deposit markdowns, in line with the data. The intuition works as follows. Consider a bad aggregate state, with a rising risk-free interest rate. In the model, this increases the households' opportunity cost of saving in deposits. Banks, therefore, exercise their market power in order to avoid a deposit flight, and raise the deposit interest rate more than proportionally relative to the risk-free rate, leading to a counter-cyclical markdown. There is, however, an interaction between market power on the asset and the liability side of banks' balance sheets. The rise in the deposit rates leads to a rise in the marginal cost. Therefore, in order to shield their profits in a recession, banks raise the credit interest rate more than proportionally relative to the marginal cost, leading to a counter-cyclical markup.

Credit market power per se amplifies aggregate fluctuations. In recessions, banks charge firms with higher interest rates. Thus, aggregate quantities (total assets, output, and consumption) all contract by more relatively to the perfect competition counterfactual. Conversely, deposit market power dampens the response of both financial and real variables. Banks, by raising markdowns in bad aggregate states, try to protect the demand for deposits, thereby allowing depositors (households) to smooth their response to shocks, preventing deposit withdrawals and the resulting contraction in lending.

Fifth, granularity plays a crucial role in shaping the business cycle in our model. Almost all the variation in aggregate output due to idiosyncratic shocks can be accounted for by shocks hitting only the *top quintile* of the banking distribution. Granular shocks are also quantitatively important relative to aggregate shocks. They account for about forty percent of the fluctuations in output due to aggregate TFP shocks. The role of granular shocks is particularly large for financial variables. Relative to an economy with only aggregate shocks, granular banking shocks account for 90 percent of the fluctuations in assets and for twice as much as the movement in net worth. Thus, in our economy, idiosyncratic shocks to large banks alone can lead to endogenous real and financial fluctuations even in the absence of any aggregate disturbances. This result complements the findings in Carvalho and Grassi (2019) for the case of non-financial firms and further advances the broader granular hypothesis agenda (Gabaix, 2011).

Finally, we leverage the dynamic nature of our framework and study properties of banking and economic crises that occur in a long simulation of the model. We employ the event-study approach that is popular in the open-economy macroeconomics literature

(Mendoza, 2010), simulate our model for a large number of periods, and identify events as incidents of aggregate output falling below a certain threshold . Our framework is good at generating realistic banking and economic crises. When the baseline Bewley Banks economy is parameterized to fit the collapse of U.S. GDP during the Great Recession, the representative-bank economy can account for less than half of the actual observed contraction of output in the data. Also, the contraction in bank assets and net worth during a banking crisis is an order of magnitude larger in the Bewley Banks economy relative to an economy with a representative bank.

We then test whether macroprudential regulation can mitigate the loss of output during crises. Doubling capital requirements improves financial stability and dampens economic contractions in typical model-simulated recessions. The reduction is, however, quantitatively small and is potentially not justified given that capital requirements generate lower *levels* of lending and production in the high-regulation steady state.

Related Literature Our paper relates to several literature strands that span macroeconomics, banking, and financial economics. The first - so-called "macro-banking" - branch embeds the financial intermediary sector into macroeconomic frameworks[1]. The broad idea can be understood as quantifying the impact of general-form financial frictions on aggregate dynamics (Kiyotaki and Moore, 1997; Bernanke et al., 1999; Cooley and Quadrini, 2001). Our paper's modelling approach is related to the seminal setup in Gertler and Kiyotaki (2010) and Gertler and Karadi (2011), which has been furthered in such works as Bocola (2016), Nuno and Thomas (2017), and Lee et al. (2020). Our study particularly emphasizes the departure from the representative intermediary and perfect banking competition assumptions and their first-order impact on business cycle fluctuations.

The present paper also builds on the vast banking literature[2]. Especially relevant to us is the growing literature on imperfect competition in banking[3]. The deposit market power channel in our framework is modelled in the spirit of Drechsler et al. (2017, 2021): via the assumption that deposits are differentiated across banking franchises and that they provide special liquidity services for households. In addition, the credit market power channel operates through imperfect substitutability of bank loans and real rigidities in the spirit of the "keeping up with the Joneses" specification (Gali, 1994). As a result, we achieve

[1]Important contributions include, among others, Cúrdia and Woodford (2001); Brunnermeier and Pedersen (2009); Adrian and Shin (2010); Jermann and Quadrini (2013); Brunnermeier and Sannikov (2014); He et al. (2016); Begenau et al. (2021); Amador and Bianchi (2021); Elenev et al. (2021); Bocola and Lorenzoni (2023).

[2]Some of the relevant studies include Diamond and Dybvig (1983); Diamond (1984); Bernanke and Blinder (1988); Bernanke and Gertler (1995); Holstrom and Tirole (1997); Allen and Gale (1998); Hellman et al. (2000); Allen and Gale (2004).

[3]Valuable contributions are Boyd and Nicolo (2005); Egan et al. (2017); Heider et al. (2019); Kurlat (2019); Drechsler et al. (2021); Polo (2021); Whited et al. (2021); Di Tella and Kurlat (2021); Wang (2022); Wang et al. (2022); Abadi et al. (2022)

heterogeneous and aggregate state-dependent markups and markdowns simultaneously.

Our paper belongs to the new, burgeoning literature on heterogeneous financial intermediaries[4]. Coimbra and Rey (2023) develop a general equilibrium model with financial intermediaries that are ex-ante heterogeneous in Value-at-Risk constraints, i.e. preferences for risk-taking. Corbae and D'Erasmo (2022) build a quantitative model of banking industry dynamics with uninsured idiosyncratic return risk and imperfect credit-market competition. Bianchi and Bigio (2022) study the credit channel of macroeconomic transmission in a macro-banking framework with stochastic deposit withdrawal shocks. Our contribution relative to this stream of papers is to embed both ex-ante and ex-post bank heterogeneity in an empirically consistent way. Moreover, we allow ex-post heterogeneity (transitory risk) to be counter-cyclical, thus generating substantial amplification of business cycles due to greater downside loan portfolio risk in recessions.

Finally, we solve our model non-linearly by building on the canonical Krusell and Smith (1996) solution method. A novel aspect of our framework is that the endogenous, dynamic distribution of bank size is a relevant state variable. We approximate the dynamics of this distribution - an infinitely dimensional object - with a small number of moments. Many of our quantitative exercises study financial and macroeconomic responses to aggregate TFP impulses. By the logic of Boppart et al. (2018), these impulse responses would be identical, as a first-order approximation, to transitional dynamics in response to zero-probability MIT shocks in a version of our model without aggregate risk. An important advantage of solving our model with full-fledged aggregate uncertainty is our ability to target and match moments of the U.S. business cycle, thus making the model more empirically-consistent, and to study the frequency and severity of banking and economic crises in its long simulation.

The rest of the paper is structured as follows. Section 2 reports stylized facts on the banking distribution in the cross-section and over the business cycle. In Section 3, we lay out our model. Section 4 describes our calibration strategy, shows the model policy functions and ergodic distributions, and demonstrates the responsiveness to aggregate fluctuations. Section 5 inspects the model mechanism by isolating each key moving part. Section 6 presents our main quantitative results and experiments. Section 7 explores the implications of heterogeneity for economic and banking crises. Finally, Section 8 concludes.

[4]Among others, the literature includes such contributions as Gerali et al. (2010); Martinez-Miera and Repullo (2010); Christiano and Ikeda (2013); Cuciniello and Signoretti (2015); Boissay et al. (2016); Jamilov (2020); Rios Rull et al. (2020); Begenau and Landvoigt (2021); Goldstein et al. (2022); Abadi et al. (2022).

2 Stylized Banking Facts

Our quantitative framework is centered around three fundamental sets of banking facts: two-sided market power, counter-cyclicality of idiosyncratic loan returns, and concentration of the distribution of size.

2.1 Two-Sided Bank Market Power

We begin by presenting evidence of market power on the asset and liability sides of bank balance sheets. Our empirical analysis leverages Call Reports data, which covers the universe of commercial banks in the United States. Our approach is similar to the lines of Corbae and D'Erasmo (2021) and De Loecker et al. (2020). Specifically, we compute quarterly measures of bank-level loan markups $\mu^k_{j,t}$ and deposit markdowns $\mu^b_{j,t}$. We define $\mu^k_{j,t}$ as a ratio of a proxy for the interest rate on loans to the marginal cost of raising a unit of credit. The markdown $\mu^b_{j,t}$ in turn is defined as the ratio of a proxy for the interest rate on risk-free asset holdings to the marginal cost of raising a unit of deposits. Our quarterly methodology replicates the yearly evidence on loan markups from Corbae and D'Erasmo (2021) and extends the analysis to markdowns which, to the best of our knowledge, is the first of its kind. Appendix A.1 provides details on the data and Appendix A.2 explains the estimation procedure.

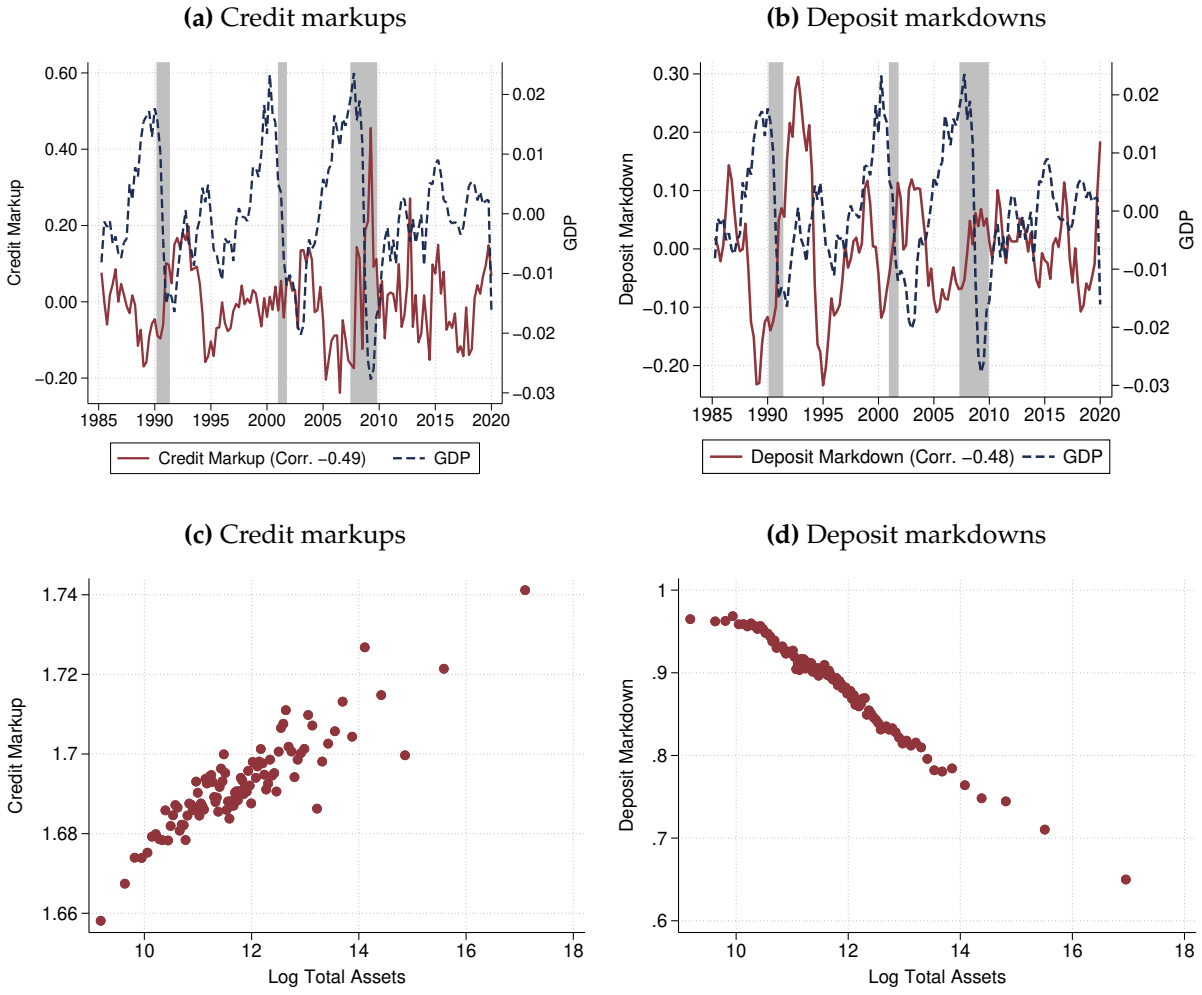
Figure 1 presents the cyclical component of bank markups and markdowns respectively in Panels (a) and (b) [5]. To arrive at these series, we construct quarterly unweighted averages for $\mu^k_t = \sum_j s_t \mu^k_{j,t}$ and $\mu^b_t = \sum_j s_t \mu^b_{j,t}$ where s_t is the inverse of the number of banks in a quarter.[6] The two series have then been logged and filtered with the Hodrick-Prescott filter with the usual smoothness parameter 1600. The same data treatment has been applied to the series of U.S. real GDP growth. We document that both credit markups and deposit markdowns are heavily counter-cyclical. The correlation of each with filtered GDP growth is around -0.49 and statistically significant at the 1% level.

The counter-cyclicality of both credit markups and deposit markdowns has implications for the pass-through from changes in marginal costs to credit and deposit interest rates. In slight anticipation of our model, suppose that business cycles are driven by a single, supply-side disturbance. Then, negative aggregate states are associated with a rising marginal cost of funds. Conditional on this, counter-cyclical credit markups imply a pass-through from the marginal cost to the credit rate that is *greater* than one-to-one. As the marginal cost rises, so does the markup, fueling a more than proportional increase in the credit rate. Counter-cyclicality of markdowns also points to a pass-through that is greater than one-to-one. The wedge between the risk-free rate and the retail rate on

[5] Appendix A.2 also shows the raw, unfiltered series.

[6] Appendix A.3 shows that results do not change if we compute weighted-average quarterly measures.

Figure 1: Bank Market Power



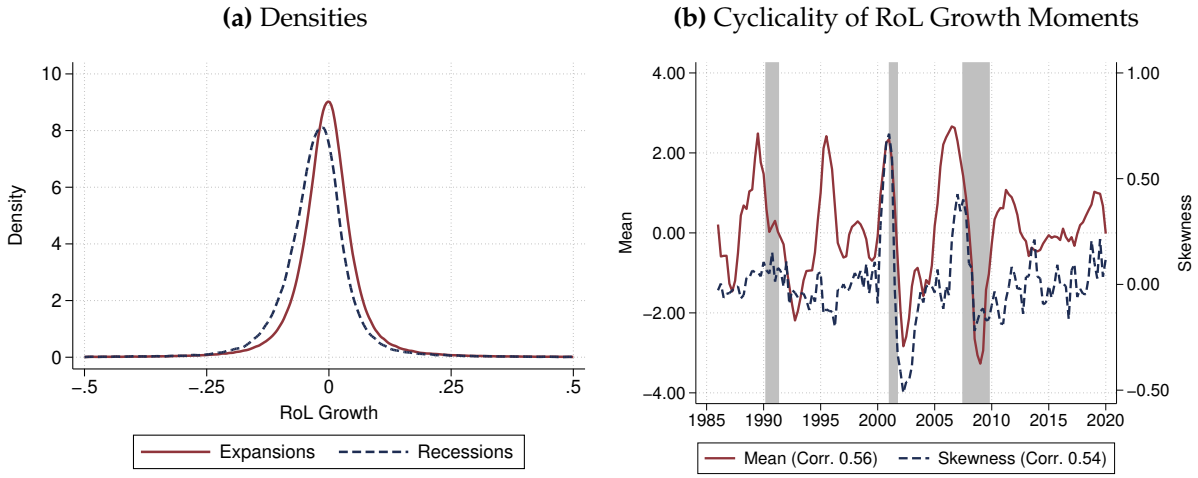
Notes:

deposits shrinks in negative aggregate states, which means that the deposit rate change is larger than the change in the risk-free rate.

Credit markup counter-cyclicality is reminiscent of the classic New Keynesian logic (Nekarda and Ramey, 2021). However, our theoretical economy will be described by a completely "real" model where real rigidities are obtained through the assumptions on technology rather than any type of nominal price stickiness. Deposit markdown counter-cyclicality reflects the notion of deposit franchise "stickiness", which has resurfaced amidst the 2023 regional banking crises in the U.S.. A key contribution of our paper is to analyze two-sided bank market power *jointly* in an empirically-motivated equilibrium setting.

Figure 1 also shows, in Panels (c) and (d), the cross-sectional relationships between markups, markdowns, and bank *size*. The panels present binned scatterplots over 100

Figure 2: Counter-Cyclical Income Risk



Notes:

equally-sized bins of (log) total assets, such that the y-axes show bin-specific unweighted averages. In addition, variables have been residualized from the time fixed effect. Market power on both sides of the balance sheet is concentrated in the right tail of size distribution: markups increase and markdowns fall with bank assets. In other words, larger banks charge simultaneously higher credit markups and lower deposit markdowns. Notice that the dual concentration of size and market power is potentially concerning from the point of view of competition regulation [7].

Before proceeding with the rest of the section, it is important to address a concern whereby our markup and markdown measures might be biased. There is a risk that the so-called "production function estimation" approach, a variant of which we follow in this paper, hinges on assumptions on the price and quantities of loan and deposit goods that could bias the outcome. Recent work by Grassi et al. (2022) for non-financial firms shows that while such approach may produce an incorrect *level* of the aggregate markup, its behavior over time - including the trend and cyclical components - is strongly correlated with the true markup measure. Thus, our cyclical components of markups and markdowns are unbiased. Furthermore, the fact that Call Reports data has exact balance sheet data on both loan and deposit returns and quantities implies, importantly, that a significant component of marginal costs is directly observable.

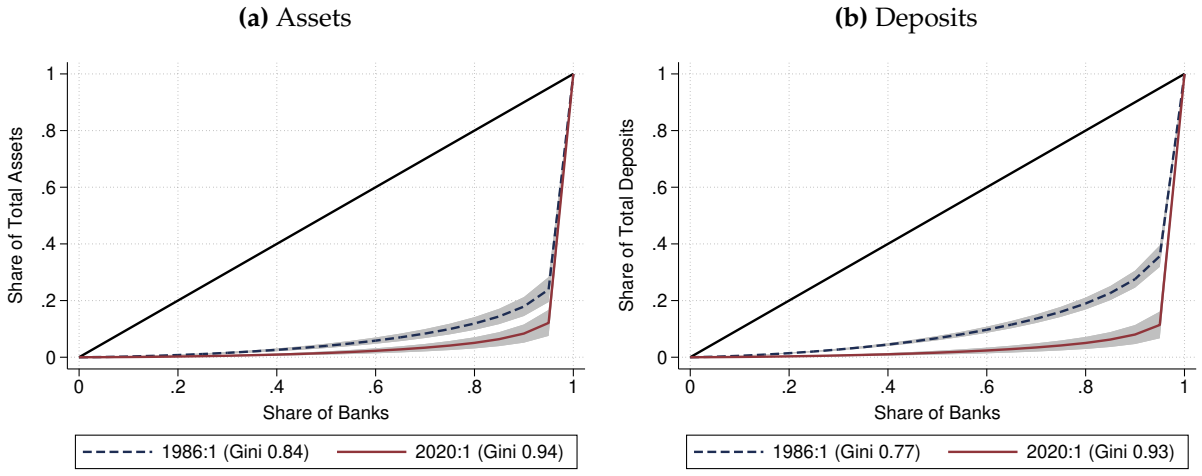
2.2 Counter-Cyclical Loan Income Risk

A second key ingredient of our quantitative analysis is the cyclicality of transitory loan return risk. A substantive literature has documented that households and firms face greater uninsured idiosyncratic risk in recessions, particularly due to the pro-cyclicality of the third moment (Bloom et al., 2019; Buch et al., 2022). A natural question to ask is whether banks face the same problem. Our measure of loan income risk is built in line with the literature (Guvenen et al., 2014). First, the realized return on loan (RoL) $r_{j,t}$ is computed as a ratio of total interest income on loans to loan holdings at the bank-quarter level. The log-difference of this variable, $\Delta r_{j,t}$, then constitutes our quarter-on-quarter transitory loan return risk measure.

Figure 2 depicts the distributions of $\Delta r_{j,t}$ for U.S. expansions and recessions, defined by the NBER criterion, and over the 1984:1-2020:1 period. The density of RoL growth is visibly pro-cyclical. The unweighted mean of RoL growth is roughly -0.3$ in expansions and -2.4% in recessions, i.e., lower by a full two percentage points. To further shed light on the cyclicality of moments of the RoL distribution, we aggregate $\Delta r_{j,t}$ to the quarterly level by computing the unweighted first, second, and third moment (statistical skewness). Each series is then HP-filtered and, additionally, smoothed with a moving-average filter with four lags. Panel (b) of the Figure plots the time-series dynamics of the resulting smoothed measures of the first and third moments. Both variables are heavily pro-cyclical, with pairwise correlations with U.S. real GDP growth equal to 0.56 and 0.54, respectively, and statistically significant at the 1% level. [8]

The pro-cyclicality of skewness of bank income risk, in particular, is a novel and important finding. It suggests that banks face greater downside risk in recessions. The implication for our modeling approach is significant: it potentially requires a departure from the standard Gaussian assumption on the distribution from which idiosyncratic return shocks are drawn. Interestingly, the second moment of the distribution of RoL growth is essentially flat over the cycle, with the correlation coefficient with GDP growth statistically indistinguishable from zero at the 10% level (not shown). This finding is consistent with the evidence in, for example, Buch et al. (2022) who show that the variance of individual-level earnings growth is acyclical in four advanced economies. All in all, pro-cyclicality of the first, a-cyclicality of the second, and pro-cyclicality of the third moment of the banks' RoL growth distribution are key motivating facts for our theoretical framework.

Figure 3: Size Concentration



Notes:

2.3 Size Concentration

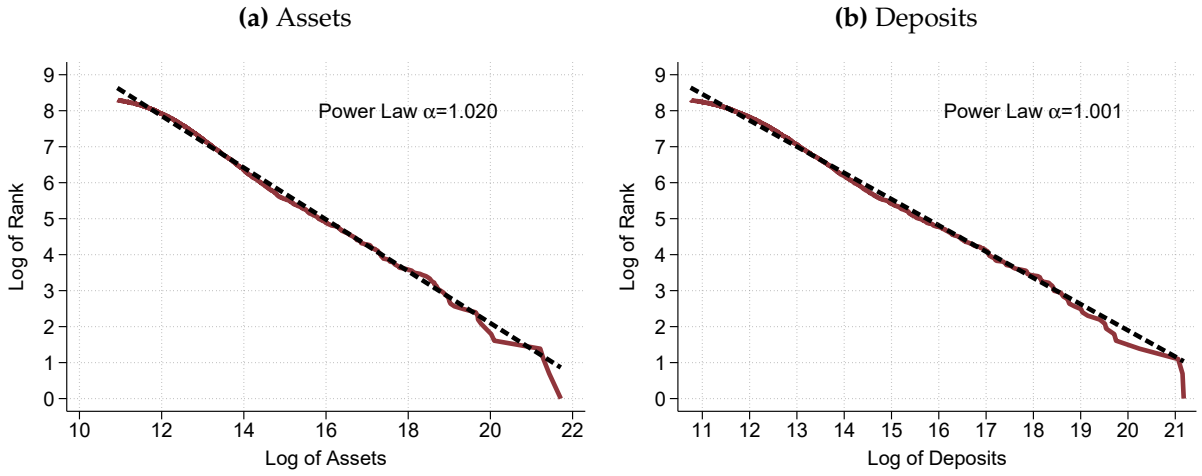
A robust feature of the U.S. banking data is the market concentration of the sector, which is not only considerably high but rising since at least the 1980s (Corbae and D'Erasmo, 2020). Figure 3 plots Lorenz curves - a standard market concentration metric - for commercial bank total assets, using Call Reports data. Departure from the equal allocation counterfactual (45-degree line) is substantial. The Gini coefficient has increased by roughly 12% over the 1986-2020 period and currently stands at 0.94. To put these numbers in further context, at present times the largest 25 banks controls roughly 95% of all assets.

Interestingly, it appears that a lesser known fact is the rise of deposit market concentration over the same period and for the same sample. Figure 3 plots the Lorenz curves for the U.S. commercial bank deposit market, also using data from Call Reports. The rise of banking concentration is even more profound when size is proxied with deposits. Consider that the deposits Gini has grown by about 20% since 1986 and is currently 0.93. What is the correct proxy for bank size is not immediately obvious. However, the 2023 U.S. regional banking crisis has put at the center stage the present discounted value of deposit franchises. Deposit franchise stickiness acts as a hedge against asset portfolio risk which could stem from, for example, rising interest rates. Therefore, deposits (especially those that are insured) act as a buffer stock against adverse fluctuations in returns on loans and other investments. Therefore, from the point of view of our quantitative model that is built around unhedged idiosyncratic rate of return risk, it could be argued that total

[7] A good measure of market power should correlate with proxies of profitability. In Appendix A.3 we show that markups and markdowns are strongly correlated with bank-level returns on assets (RoA)

[8] Appendix A.3 shows that our results hold at the bank holding level of aggregation.

Figure 4: Granular Banks



Notes:

deposit holding is a more relevant notion of size.

2.4 Granularity

The seminal contribution by Gabaix (2011) has put forth the so-called "Granular Hypothesis": idiosyncratic firm-level disturbances by themselves could theoretically be enough to generate aggregate fluctuations for as long as some firms are abnormally large, i.e., granular. The granular hypothesis has been investigated extensively in the contexts of non-financial firms (Carvalho and Grassi, 2019), bank portfolio concentration (Galaasen et al., 2020), trade (Gaubert and Itskhoki, 2021), and exchange rates (Camanho et al., 2022). In the context of the U.S. banking sector, we have already documented the extreme degrees of size heterogeneity in the markets of both asset and deposit holdings. We now examine whether the granular hypothesis applies, thus making the right tails of the asset and deposit distributions relevant from a macroeconomic perspective.

There are two simple tests of granularity that are typically run in the literature. First, the ranks-size rule compares log-size against log-rank (Gabaix, 2009). A special case of this comparison is the famous Zipf's law, which arises if the relationship is approximated with a straight negatively-sloped line. We run this test for the U.S. banking sector, restricting the sample to 2020:1 and the largest 4,000 banks. Figure 4 plots log-ranks of assets (Panel A) and deposits (Panel B) on the y-axis against log-size on the x-axis. The striking result is how tightly straight lines can summarize the data. The R^2 of linear regressions of log-rank on log-size, on both panels, is above 0.99.

The second test is understood in distributional terms: consider $\Pr(\text{Size} > S) = \kappa/S^{\alpha}$

which means that the likelihood of a bank being greater than some S is proportional to α/S. We estimate α with maximum likelihood methods for both assets and deposits, again for 2020:1 and for the largest 4,000 banks, and obtain the values of 1.02 and 1.001 for assets and deposits, respectively. First of all, this implies that Zipf's law is in fact a good first-order approximation of the data. Second, there is strong evidence in support of the granular hypothesis applying to financial intermediaries. In other words, idiosyncratic shocks to large banks alone could generate non-diversified grains of financial activity and drive aggregate fluctuations. We return to this important point in the modelling sections of the paper.

Summary To conclude, a realistic quantitative macro-banking model should feature several empirically verified modeling devices. First and foremost, there must exist a well-defined notion of bank size - measured by either assets, deposits, or net worth (capital). Second, the model needs to deliver a reasonable degree of size concentration with a non-trivial share of very large institutions. Third, idiosyncratic shocks to those large institutions should be enough to generate reasonable aggregate fluctuations. Fourth, competition should be imperfect on both the asset and the deposit side of the balance sheet; banks should be allowed to charge markups to borrowers and markdowns to depositors in a flexible, heterogeneous fashion. Finally, banks should be exposed to idiosyncratic shocks that get worse in negative aggregate states. This counter-cyclicality should be driven, at least in part, by the pro-cyclicality of the third moment of the density of risk.

3 Model

This section lays out a business cycle model with heterogeneous banks. Time is discrete and infinite. The economy is populated by four agents: a representative household, a representative firm that produces the capital good, a final good producer, and a continuum of heterogeneous financial intermediaries (banks, for short) indexed by $j \in [0, 1]$.

3.1 Technology

There is a continuum of measure one 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:

$$Y_t = A_t K_t^{\alpha} L_t^{1-\alpha}, \quad 0 < \alpha < 1 \tag{1}$$

Output is transformed into future capital or current consumption. Capital accumulates

over time according to the law of motion:

$$K_{t+1} = (1 - \delta)K_t + I_t \tag{2}$$

where $0 \leq \delta \leq 1$ is the rate of depreciation and I_t is aggregate investment. Returns on aggregate capital holdings can be represented by:

$$R^k_{t+1} = \frac{A_{t+1}\alpha K^{\alpha-1}_{t+1} + (1 - \delta)Q_{t+1}}{Q_t} \tag{3}$$

where Q_t is the price of capital determined by the capital production block. Aggregate productivity A_t is stochastic and takes on two possible values: A^h and A^l which represent, respectively, good and bad aggregate states. The shock follows a first-order Markov structure with $\pi_{\mathbf{a}}$ the matrix of transition probabilities.

3.2 Firms

Capital good producers are cash-strapped and require bank financing in the form of equity-type claims, which we will be referring to interchangeably as assets or credit. In exchange for bank credit, firms pledge the future realized return on the aggregate capital stock (3). We assume that firms possess a technology to costlessly convert units of credit into differentiated units of capital, which get immediately aggregated. The credit market is imperfectly competitive and aggregate capital K_t is assembled by the following aggregator:

$$K_t = \int_0^1 \left[\left(k_t(j) - \gamma_1 K_t^{\gamma_2} \right)^{\frac{\theta_k - 1}{\theta_k}} dj \right]^{\frac{\theta_k}{\theta_k - 1}} \tag{4}$$

where $\theta_k > 1$ is the elasticity of substitution and $\{\gamma_1, \gamma_2\}$ govern the degree of production externalities. Firms solve a cost minimization problem, which yields the following demand curve for differentiated units of capital:

$$k_t(j) = \left(\frac{q_t(j)}{Q_t} \right)^{-\theta_k} K_t + \gamma_1 K_t^{\gamma_2} \tag{5}$$

where Q_t is the aggregate credit rate. The asset demand elasticity can be shown to equal: $\theta_k \frac{k_t(j) - \gamma_1 K_t^{\gamma_2}}{k_t(j)}$. We can now write down the Lerner condition for the credit market, which decomposes $q_t(j)$ in terms of the markup $\mu^k_t(j)$ and the marginal cost $MC_t(j)$, to be

determined later:

$$q_t(j) = \underbrace{\frac{\theta_k \frac{k_t(j) - \gamma_1 K_t^{\gamma_2}}{k_t(j)}}{\theta_k \frac{k_t(j) - \gamma_1 K_t^{\gamma_2}}{k_t(j)} - 1}}_{\text{Credit markup } \mu_t^k(j)} \underbrace{MC_t(j)}_{\text{Marginal Cost}} \tag{6}$$

The capital aggregator (4) requires a special discussion. It is an amalgamation of several directions and concepts in the literature. First, canonical Pollak (1971) demand is nested as a special case whenever $\{\gamma_1 > 0, \gamma_2 = 0\}$. This case yields an equilibrium markup which is greater than unity but homogeneous across banks. Second, under the $\{\gamma_1 < 0, \gamma_2 > 0\}$ parametrization we obtain the classic "keep up with the Joneses" specification, or a type of *positive* production externalities (Gali, 1994). In this case, increases in the aggregate production of capital in the economy raise the private marginal product, forcing an individual firm to produce more capital for which, in turn, it requires more funding from the banks. This specification would yield a distribution of markups that, correctly, increase with relative bank size. Third, a variant of the classic habit formulation is obtained whenever $\{\gamma_1 > 0, \gamma_2 = 1\}$ (Ravn et al., 2006). Finally, the constant elasticity of substitution case obtains for $\gamma_1 = 1$. The diagram below summarizes all the possible situations:

$$\gamma_1 \begin{cases} = 0 & \text{CES} \\ > 0 & \text{Negative production externalities} \\ < 0 & \text{Keeping up with the Joneses} \end{cases} \qquad \gamma_2 \begin{cases} = 0 & \text{Pollak demand} \\ = 1 & \text{Linearity} \\ > 1 & \text{Convexity} \end{cases}$$

In order to match the observable cross-sectional distribution and cyclicality of loan markups in the data, we first require $\gamma_1 < 0$, i.e., the keeping up with the Joneses specification. This grants us the correct relationship between size and markups in the *cross section*. In addition, we need $\gamma_2 > 1$, i.e., convexity. General equilibrium adjustment of the full distribution of claims $\int k(j)$ in response to an aggregate shock nullifies private, bank-level partial-equilibrium markup changes in a way that makes the aggregate markup a-cyclical. In other words, while private markups $\mu^k(j)$ rise when private relative size is high after a positive shock, in equilibrium the aggregate K adjusts and the aggregate markup remains unchanged. With $\gamma_2 > 1$, we are able to obtain a pro-cyclical aggregate demand elasticity. In response to a positive aggregate productivity shock, once aggregate capital K_t adjusts, the convexity of the production function kicks in and the aggregate demand elasticity increases endogenously due to second-round effects, thus lowering the aggregate markup. This grants the correct unconditional correlation between production and markups along the *time series*. All in all, in order to match two moments in the data - one cross-sectional and one time-series - we require two parametric restrictions: $\gamma_1 < 0$ and $\gamma_2 > 1$.

In models with aggregate uncertainty, like ours, parsimony is crucial. Our aggregator

is designed to be operational and flexible. At the same time, it is a reduced-form representation of much more complex, deeper mechanisms at work. Consider the canonical contribution of Chevalier and Scharfstein (1996) who develop a theory of counter-cyclical markups and capital-market frictions. In their theory, firms face an occasionally-binding liquidity constraint that binds in bad aggregate states. Illiquidity forces firms to increase profits in the short run by raising prices. Our capital production function is entirely consistent with this economic mechanism. Recessions are associated with lower realized returns for the financial intermediaries. Banks, in order to prevent a complete depletion of net worth, have an incentive to sacrifice private market share in the short run and improve profitability by raising markups.

3.3 Households

The representative household's preferences are separable intertemporally and discounted at the rate of $\beta \in (0,1)$. The household derives utility from consumption and bank deposit holdings as well as disutility from labor. The household can save in the form of one-period deposits or mutual funds. To motivate imperfect competition in the market for bank deposits, we assume that deposits provide special liquidity services, similarly to the setup of Drechsler et al. (2017, 2021) and Bellifemine et al. (2022) or more generally to the money-in-utility framework (Sidrauski, 1967; Gali, 2008; Walsh, 2010). Mutual funds are risk-less investments but provide no liquidity utility. Both vehicles pay guaranteed, state non-contingent rates of returns. Flow utility, which features non-separability between consumption and hours in the spirit of Greenwood et al. (1988) and separability with respect to deposit holdings, takes the form of:

$$U(C_t, L_t, B_t) = \frac{1}{1-\phi}\left(C_t - \chi_1 \frac{L_t^{1+\chi_2}}{1+\chi_2}\right)^{1-\phi} + \nu_1 \frac{B_t^{1-\nu_2}}{1-\nu_2} \tag{7}$$

where $\frac{1}{\chi_2}$ is the elasticity of labor supply, $\frac{1}{\nu_2}$ is the elasticity of deposit supply, χ_1 gauges labor disutility, and $\frac{1}{\phi}$ is the intertemporal elasticity of substitution. Deposit products are imperfect substitutes across banking franchises, indexed by j, and assembled into the aggregate stock of deposits by the following aggregator:

$$B_t = \left[\int_0^1 b_t(j)^{\frac{\theta_b+1}{\theta_b}} dj\right]^{\frac{\theta_b}{\theta_b+1}} \tag{8}$$

with $\theta_b > 0$ the elasticity of substitution across deposit franchises. The flow budget

constraint is given by:

$$C_t + \int_0^1 b_t(j)dj + M_t \leq R_t M_{t-1} + \int_0^1 R_t^b(j) b_{t-1}(j) + L_t W_t + \text{Div}_t + T_t \tag{9}$$

where M_t are mutual fund holdings, W_t is the competitive wage rate, $R_t^b(j)$ is the non-contingent bank-specific interest rate on deposits to be determined in equilibrium, R_t is the real risk-free interest rate, T_t are lump-sum taxes, and Div_t are lump-sum transfers of bank dividends.

The first-order condition with respect to $b_t(j)$ yields a Lerner-type formula for deposit interest rates:

$$R_{t+1}^b(j) = R_{t+1}\left(1 - \left[\underbrace{\frac{U_{B,t}\left(C_t, L_t, B_t\right)}{U_{C,t}\left(C_t, L_t, B_t\right)}}_{\text{Marginal Liquidity Preferences}} \overbrace{\left(\frac{b_t(j)}{B_t}\right)^{\frac{1}{\theta_b}}}^{\text{Product Differentiation}}\right]\right) \tag{10}$$

where $U_{B,t}$ and $U_{C,t}$ denote marginal utility operators and $R_{t+1} = \left[\beta \mathbb{E}_t \frac{U_{C,t+1}(C_{t+1}, L_{t+1}, B_{t+1})}{U_{C,t}(C_t, L_t, B_t)}\right]^{-1}$ is the risk-free interest rate, which is pinned down by a first-order condition with respect to risk-less mutual fund holdings. The dynamic, heterogeneous deposit markdown can be defined as follows:

$$\mu_t^b(j) = \frac{R_{t+1}^b(j)}{R_{t+1}} \leq 1$$

The deposit market power of banks is determined by two factors. First, the "marginal liquidity preferences" term which is governed by the cyclicality of the marginal utility of deposit holdings. In recessions, the marginal utility of consumption rises. If the marginal utility of deposit holdings rises but by less, such that the ratio is pro-cyclical, then the markdown is counter-cyclical - as in the data. In other words, the relative marginal benefit from deposits needs to fall in response to negative aggregate shocks. Banks, internalizing this effect, attempt to protect the deposit franchise by raising the markdown, i.e., by shrinking the deposit spread and prioritizing market share retention in the short run. Second, the "product differentiation" term, which can also be understood as the degree of substitutability across deposit franchises. It is evident from (B2) that the model can nest (a) perfect deposit market competition (if ν_1=0) or (b) a homogeneous markdown (if $\theta_b \to \infty$ for some finite $\nu_1 > 0$). In the baseline economy, the markdown is time-varying, heterogeneous, and proportional to relative deposit size.

Finally, the first-order-condition with respect to labor supply is standard:

$$L_t = \left(\frac{W_t}{\chi_1}\right)^{-\frac{1}{\chi_2}} \tag{11}$$

3.4 Financial Intermediaries

The financial intermediation sector is populated by a continuum of measure one banks which are indexed by $j \in [0,1]$. Banks' role in the economy is to source deposits $d_t(j)$ from households and invest into the capital producing firms via claims $k_t(j)$. Deposits pay a state non-contingent interest rate $R_t^b(j)$. Banks accumulate net worth $n_t(j)$, maximize the present discounted value of their franchise value, exit the economy with a probability $1-\sigma$, upon which their entire franchise gets transferred to the household in the form of dividends.

Ex-ante Heterogeneity Banks are ex-ante non-identical due to permanent differences in the efficiency of intermediation $\kappa \in \Theta$. A higher $\kappa(j)$ allows banks to consistently identify more profitable lending opportunities, yielding permanently higher returns for the same amount of claims held. For example, this could be due to differences in monitoring skill and ability (Diamond, 1984). In addition, markets are incomplete and portfolio returns feature uninsured idiosyncratic risk $\xi_t(j)$. All banks earn a common aggregate return on capital R_t^k, which is perturbed by the permanent and transitory components of rate of return heterogeneity. Bank-level total portfolio return $R_t^T(j)$ can thus be written as:[9]

$$R_t^T(j) = \kappa(j)\xi_t(j)R_t^k \tag{12}$$

At the beginning of time, permanent return types $\kappa(j)$ are drawn by nature from a power law distribution, specifically the Type 1 Pareto density with the shape parameter $\alpha > 0$ and a normalizing minimal value κ_m:

$$\kappa(j) \sim P(\alpha, \kappa_m), \quad Prob(\kappa > \kappa_m) = \left(\frac{\kappa_m}{\kappa}\right)^{\alpha} \tag{13}$$

As the shape parameter approaches unity, the departure from standard laws of large numbers becomes more severe, and the "Granular Hypothesis" becomes more relevant (Gabaix, 2011). Specifically, in the $\alpha \in [1,2)$ region, idiosyncratic shocks to individual banks that operate in the right tail of the distribution of permanent income are more likely to not wash out in the aggregate. This insight will be important in **Section XXX** when we discuss the role of large banks in aggregate financial and business cycles. This comes

[9]The permanent-transitory risk mixture is common to other work in the literature. See, e.g., Guvenen et al. (2023) in the context of household income risk and wealth taxation.

in addition to enabling the model to generate a realistic degree of size concentration. In equilibrium, banks with an abnormally large $\kappa(j)$ are much larger than the median bank, raising the model-implied GINI coefficient.

Ex-post Heterogeneity Transitory rate of return risk $\xi \in \Xi$ follows an AR(1) process with shocks $e_t(j)$ drawn from a non-Gaussian, aggregate state-dependent distribution. Specifically, we employ the Hansen (1994) Skewed-t density with time-varying parameters $\mu_{\epsilon,t}$, $\lambda_{\epsilon,t} \in (-1, 1)$, and $\eta_t \in (2, \infty)$ which, respectively, govern the mean, skewness, and degrees of freedom of the distribution. Formally:[10]

$$\xi_t(j) = (1 - \rho_\xi)\mu_\xi + \rho_\xi \xi_{t-1}(j) + \epsilon_t(j) \tag{14}$$

and

$$\epsilon_t(j) \sim g(z_t(j)|\eta_t, \lambda_{\epsilon,t}) = \begin{cases} bc\left(1 + \frac{1}{\eta_t - 2}\left(\frac{bz_t(j)+a}{1-\lambda_{\epsilon,t}}\right)^2\right)^{-(\eta_t+1)/2}, & z_t(j) < -a/b \\ bc\left(1 + \frac{1}{\eta_t - 2}\left(\frac{bz_t(j)+a}{1+\lambda_{\epsilon,t}}\right)^2\right)^{-(\eta_t+1)/2}, & z_t(j) \geq -a/b \end{cases} \tag{15}$$

with $z_t(j)$ a variable drawn from the standard Normal distribution and triad $\{a, b, c\}$ known constants. Following Hansen (1994), in order to control the number of free parameters, we impose an exact mapping from $\lambda_{\epsilon,t}$ onto η_t: $\eta_t = L + \frac{U-L}{1+\exp(-\lambda_{\epsilon,t})}$ such that the only source of time-varying deviation from non-normality is the skewness parameter $\lambda_{\epsilon,t}$. Figure 5 illustrates how Hansen's skew-t departs from the Gaussian density. The major difference is the introduction of a sharp, prolonged left-tail. Left-skewness appears whenever $\lambda_{\epsilon,t} < 0$.

Balance Sheet Banks start each period with an initial stock of net worth $n \in \mathbf{N} \subset \mathbf{R}_+$. Each banking franchise is required to pay operational, non-interest variable expenses that are increasing and convex in book assets under management: $\zeta_1 k_t^{\zeta_2}$ with $\zeta_1 > 0, \zeta_2 > 1$. Alternatively, private adjustment of the quantity of loans is costly and the cost must be incurred at the franchise level. Convexity of these costs breaks scale invariance and makes bank size (net worth) a relevant state variable. The law of motion of bank-level net worth $n_t(j)$ can thus be written as:

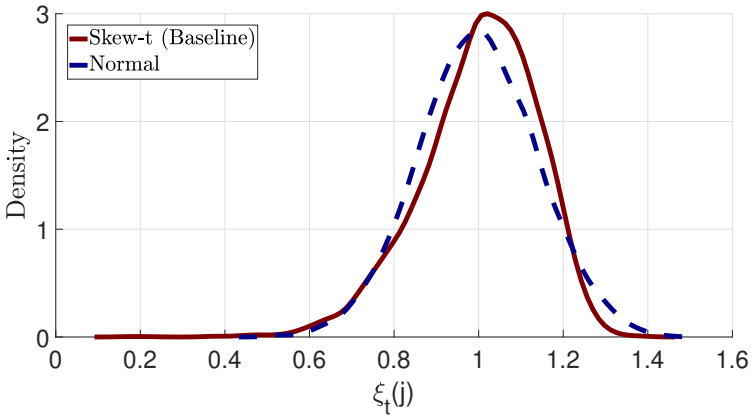
$$n_{t+1}(j) = R^T_{t+1}(j)q_t(j)k_t(j) - R^b_{t+1}(j)b_t(j) - \zeta_1 k_t(j)^{\zeta_2} \tag{16}$$

At all times and for all banks, the balance sheet constraint must hold:

$$b_t(j) + n_t(j) = k_t(j) \tag{17}$$

[10]Our approach complements other existing methods in the literature. For example, see **?** who model $\xi_t(j)$ as a random variable that is drawn from a mixture of several normally distributed random variables.

Figure 5: Hansen's Skew-t Density



Notes: Skewed-t density (solid) vs Normal density (dashed).

To motivate some role for macroprudential policy, and in anticipation of our quantitative experiment that involves raising capital requirements, we follow Gertler and Kiyotaki (2010) and Gertler and Karadi (2011) and postulate that the banking sector is subject to an agency friction. Bankers have an incentive to divert a fraction λ of the franchise $V_t(j)$ for personal use. If a diversion is successful, the franchise is bankrupt and only the remaining fraction $1-\lambda$ is recovered by depositors. We assume that the household sector is unable to impose a ceiling on bank asset growth. However, the government can. In particular, λ is set such that it exactly offsets the bankers' diversion incentive and, in equilibrium, this yields the following constraint on leverage:

$$\lambda k_t(j) \leq V_t(j) \tag{18}$$

The presence of market incompleteness and uninsured idiosyncratic rate of return risk implies that there is an inherent probability of fundamental bank insolvency. This occurs whenever the draw $\xi_t(j)$ is sufficiently low such that $n_t(j)$ is drawn down to zero, below which the bank cannot operate. Insolvency risk is priced competitively into the distribution of the market price of deposits $R_t^b(j)$. We assume that the government operates a deposit insurance scheme - an essential pillar of traditional banking (Farhi and Tirole, 2020). The scheme neutralizes the insolvency risk spread such that, in the absence of any markdown distortions, the deposit rate equals the non-contingent risk-free rate R_t. The scheme is financed with lump-sum non-distortionary taxation of the household sector.

Marginal Cost It can be shown that the relevant marginal cost for the dynamic banking problem is:

$$MC_t(j) = \frac{R_t^b + \zeta_1 \zeta_2 k_t(j)^{\zeta_2 - 1} + \tilde{\lambda}_t(j)\lambda}{\mathbb{E}_t R_{t+1}^T(j)} \tag{19}$$

where $\tilde{\lambda}_t(j)$ is the Lagrange multiplier on the leverage constraint. The total marginal cost is increasing in marginal interest and non-interest costs of producing a single unit of bank credit, the shadow cost of the incentive constraint, and decreasing in expected portfolio profitability. The latter term, in the spirit of Melitz (2003), essentially acts as an "efficiency" shifter, making the franchises of the more profitable banks less costly to operate.

Dynamic Problem Henceforth, we adopt a recursive notation and, for ease of notation, we drop the time and (j) indexations temporarily. The aggregate state of the economy $\mathbf{S}$ is characterized by a dyad $\{\mu, A\}$. μ is a probability measure defined on the Borel algebra B that is generated by the open subsets of the product space $\mathbf{B} = \mathbf{N} \times \mathbf{\Theta} \times \mathbf{\Xi}$, representing the endogenous, time-varying joint cross-sectional distribution of bank net worth, permanent return types, and transitory return draws. The law of motion of the distribution is denoted by $\mathbf{\Gamma}$ such that $\mu' = \mathbf{\Gamma}(\mu(\mathrm{n}, \kappa, \xi), \mathrm{A}, \mathrm{A}')$. Note that A′ is included in the law of motion because idiosyncratic risk is aggregate state-dependent. The law of motion that concerns A is exogenous and described by $\pi_{\mathbf{a}}$.

The relevant idiosyncratic state vector $\mathbf{s}$ includes net worth n as well as the permanent and transitory components of return heterogeneity κ and ξ. The stream of future flows of net worth is discounted with the marginal rate of substitution $\Lambda(\mathbf{S'}, \mathbf{S})$, which is pinned down by the household problem, and augmented by the exogenous death rate $1-\sigma$. Banks take their initial net worth as given and choose how much to lend to firms k and at what rate q, and how much to borrow from households b and at what rate R^b. Competitive structures in the credit and deposit markets are taken as given. The problem takes the following form:

$$\max_{\{k,q,b,R^b\}} V(\mathbf{s}; \mathbf{S}) = \mathbb{E}_{\mathbf{s},\mathbf{S}}\{\Lambda'(\mathbf{S'}, \mathbf{S}) \left[(1-\sigma)n' + \sigma V'(\mathbf{s'}; \mathbf{S'})\right]\} \tag{20}$$

subject to:

$$
\begin{aligned}
n' &= R^{T'}(\mathbf{s'};\mathbf{S'})qk - R^{b'}b - \zeta_1 k^{\zeta_2} \\
b + n &= k \\
R^T(\mathbf{s};\mathbf{S}) &= \kappa\xi R^k(\mathbf{S}) \\
\lambda k &\leq V(\mathbf{s};\mathbf{S}) \\
R^{b'} &= R'(\mathbf{S})\left(1 - \left[\frac{U_B(\mathbf{S})}{U_C(\mathbf{S})}\left(\frac{b}{B(\mathbf{S})}\right)^{\frac{1}{\theta_b}}\right]\right) \\
q &= \frac{\theta_k \frac{k-\gamma_1 K(\mathbf{S})^{\gamma_2}}{k}}{\theta_k \frac{k-\gamma_1 K(\mathbf{S})^{\gamma_2}}{k} - 1}\frac{R^b + \zeta_1\zeta_2 k^{\zeta_2 - 1} + \tilde{\lambda}(\mathbf{s};\mathbf{S})\lambda}{\mathbf{E}_S R^{T'}(\mathbf{s'};\mathbf{S'})} \\
\mu' &= \mathbf{\Gamma}(\mu, \mathrm{A}')
\end{aligned}
$$

as well as the exogenous laws of motion for ξ and A.

3.5 Market Clearing and Equilibrium

Clearing of the market for firm claims implies that the distribution of bank credit aggregates and equals to the aggregate demand for capital from firms:

$$K'(\mathbf{S}) = \int_{\mathbf{B}} k^*(\mathbf{s};\mathbf{S})\mu(n,\kappa,\xi)d\xi d\kappa dn \tag{21}$$

where $k^*(\mathbf{s};\mathbf{S})$ denotes the banks' policy function for firm claims. Similarly, deposit market clearing requires:

$$B(\mathbf{S}) = \int_{\mathbf{B}} b^*(\mathbf{s};\mathbf{S})\mu(n,\kappa,\xi)d\xi d\kappa dn \tag{22}$$

where $b^*(\mathbf{s};\mathbf{S})$ denotes the banks' policy function for deposit demand.

Goods market clearing requires:

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

Finally, optimal choices of credit and deposit rates $\{q^*, R^{b*}\}$ are appropriately summed up to, respectively, the aggregate credit rate Q and deposit rate R^b.

Equilibrium A recursive competitive equilibrium is the law of motion of the banking distribution Γ, the bank value function V, policy functions for banks $\{k^*, b^*, q^*, R^{b*}\}$, and policy functions for the household $\{c^*, l^*, b^*\}$ such that, given a vector of aggregate pricing functions $\{R, R^k, Q, W\}$, (i) value and policy functions of all agents solve the corresponding

decision problems; (ii) Γ is consistent with agents' optimization; (iii) all markets clear.

3.6 Numerical Methods

MAY BE PUT IN APPENDIX?

In order to solve our model, we resort to non-linear computational methods. Our algorithm builds heavily on the seminal contribution of Krusell and Smith (1998). A key numerical challenge is that μ - an infinitely-dimensional object - is an endogenous state variable in the model. More specifically, in order to make private quantity and pricing decisions at time t, each bank needs to have an estimate of the return on capital in $t+1$, which in turn depends on the full distribution of future lending decisions $\int k_{t+1}(j)$. To bypass the curse of dimensionality, we assume that banks form linear, limited-information forecasts based on a small set of $\mathbf{K}$ moments of the distribution, denoted as $\mathbf{m} \equiv (\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_K)$. The law of motion of the distribution can then be re-formulated in terms of the dynamics of its moments: $\mathbf{m}' = \Gamma(\mathbf{m}, A, A')$.

The algorithm consists of three basic steps. First, conditional on some initial guess for Γ^i and the pricing functions, each agent's problem is solved using standard non-linear methods, yielding the first set of candidate value and policy functions $\{V^*, k^*, b^*_j, q^*, R^{b*}, c^*, l^*, b^*\}^i$. We solve the dynamic problems on a sparse grid of idiosyncratic and aggregate state variables. To this end, we discretize both ξ and κ and use interpolation to compute values of policy and value functions not on the grid. Second, conditional on the just-computed policy functions, we simulate a panel of $\mathbf{I}$ banks that runs for $\mathbf{T}$ periods. Third, we run a linear regression on the simulated data and obtain a new candidate Γ^{i+1}. Algorithm stops as soon as we achieve convergence of both the outer and the inner loops.

Our numerical approach, while transparent and efficient, has two possible limitations. First, in our setup we have exogenously imposed both the number and the set of moments that banks use to form expectations of future returns. In practice, endogenous information acquisition may incentivize different agents to acquire different magnitudes and intensities of information, potentially in an aggregate state-dependent manner (Broer et al., 2022a,b). Second, it is possible to impose more general, non-linear limited information forecasts of the aggregate state μ, specifically by leveraging neural networks (Nuno et al., 2023). However, we do not pursue these complications because our solution is already accurate.

4 Taking the Model to the Data

This section first describes how we take our model to the data and proceeds by analyzing some of its dynamic and cross-sectional equilibrium properties.

4.1 Calibration

One period corresponds to a quarter. Table 1 reports calibration targets that we use to pin down certain parameters internally. It also lists several additional business-cycle and steady-state moments. Table 2 summarizes the values we have assigned to every model parameter.

Technology, Firms, and Households We begin with the technology parameters. Aggregate productivity can take two values: $A^h = 1.0045$ and $A^l = 0.9955$. These values are chosen in order match the empirical volatility of U.S. real GDP growth. The transition probability matrix is set to $\pi_{H,L} = \begin{Bmatrix} 0.87 & 0.13 \\ 0.13 & 0.87 \end{Bmatrix}$ in order to deliver an autocorrelation coefficient of roughly 0.95 (Smets and Wouters, 2007).

The firms' block consists of three key parameters. First, we set θ_k to 2.05 in order to target the empirical asset-weighted average credit markup of 1.7. Second, the linear term γ_1 in the capital aggregation equation (4) is set to -0.02 in order to target the cross-sectional elasticity of credit markups with respect to bank size, which is 0.0048 in the data. To arrive at this value empirically, we first clean (log) markups from the bank-specific averages, thus discarding the permanent component of markup heterogeneity. Second, we run a panel linear regression of the de-meaned (log) markup on (log) real total bank assets and a time fixed effect. Third, γ_2 is assumed to equal 2, which helps deliver pro-cyclical movements in the aggregate demand elasticity.

We now move on to the household block. The discount factor is fixed at $\beta = 0.996$ in order to target an annualized risk-free rate of roughly 1.6%. Relative risk aversion ϕ is set to 2, a standard value in the literature. Labor disutility, χ_1, is calibrated in order to exactly deliver steady-state hours of 0.3. In other words, the household works 30% of the time. χ_2 is set to unity as is common in macroeconomic studies (Nuno and Thomas, 2017; Kaplan et al., 2018). θ_b is calibrated in order to target the empirical mean of the deposit markdown series. In the data, this value is roughly 0.8 and is obtained by taking an asset-weighted average of estimated markdown values. ν_1 is calibrated in order to target the cross-sectional elasticity of deposit markdowns with respect to bank size. The empirical procedure is identical to the case of loan markups - detailed above. The estimated elasticity in the data is -0.0050. Finally, ν_2 is set to 0.1 in order to deliver an elasticity of deposit supply of ten and a sufficiently pro-cyclical ratio of relative marginal liquidity preferences and, by extension, a counter-cyclical aggregate deposit markdown.

Banks The banking block is comprised of several standard and some novel parameters. The bank survival rate is set to 0.9, yielding an expected life duration of 10 years, which is in the range of values commonly used in the literature (Gertler et al., 2016, 2020; Lee

Table 1: Calibration Targets and Moments

Steady State		
Moment	Data	Model
Hours worked (target)	0.3	0.3
Average loan markup (target)	1.7	1.7
Average deposit markdown (target)	0.8	0.8
Commercial bank leverage (target)	6.5	6.5
Loan markup - size elasticity (target)	0.0048	0.0052
Deposit markdown - size elasticity (target)	-0.0050	-0.0049
Commercial bank assets GINI	0.94	0.70
Commercial bank deposits GINI	0.93	0.73
Financial and Business Cycles		
Moment	Data	Model
σ_Y (target)	0.011	0.011
σ_C/σ_Y (target)	0.62	0.62
$\rho_{C,Y}$	0.65	0.97
$\rho_{L,Y}$	0.86	0.92
$\rho_{K,Y}$	0.38	0.96
$\rho_{N,Y}$	0.15	0.64
$\rho_{LEV,Y}$	0.24	0.64
$\rho_{\mu^k,Y}$	-0.49	-0.92
$\rho_{\mu^b,Y}$	-0.48	-0.78

et al., 2020). The fraction of divertible assets λ equals 0.12, a value which approximates aggregate capital requirements in most countries.

A key parameter that drives permanent bank returns inequality in the model is the Pareto shape index α. Motivated by our empirical exercises, we assume that the Zipf's law holds and set α to unity. To calibrate the persistence ρ_ξ of transitory return risk, we adopt two distinct strategies. First, following Bellifemine et al. (2022) closely we bring our return process (12) to the data by estimating a linear panel fixed effects model with AR(1) disturbances in the spirit of Baltagi and Wu (1999). Our main variable of interest is the return on loans (RoL), defined as the ratio of interest income on loans over total loans. We remove the aggregate component by subtracting from the returns their quarterly averages. We then run a linear regression of de-meaned returns on a bank fixed and an AR(1) component, the latter being estimated with the Durbin-Watson estimator. We find that $\hat{\rho}_\xi$ is 0.526. Our second approach relies on estimates in Galaasen et al. (2020) who use Norwegian bank-firm matched loan-level data and report persistence parameters - at various levels of aggregation - for an uninsured idiosyncratic loan risk process. At the levels of a borrower or a banking franchise, estimates range from 0.1 to 0.32. All in all, evidence points to transitory loan return risk of banks being highly non-persistent. In the model, we set ρ_ξ to 0.5. The volatility of return risk σ_ξ is set to 0.12, which corresponds to

Table 2: Model Parametrization

Parameter	Value	Description	Parameter	Value	Description
		Technology			**Banks**
σ_a	0.45	Standard deviation, aggregate risk (%)	σ	0.9	Bank survival rate
$\pi_{L,L}$,$\pi_{H,H}$	0.87	Transition probability, aggregate risk	λ	0.12	Fraction of divertible assets
		Firms	α	1	Permanent returns, Pareto shape
γ_1	-0.02	Capital aggregator, linear	ρ_ϵ	0.5	Transitory returns, persistence
γ_2	2	Capital aggregator, quadratic	σ_ϵ	0.12	Transitory returns, standard deviation
θ_k	2.05	CES markup	μ_H	0	Mean of transitory returns, high state
		Households	μ_L	-0.02	Mean of transitory returns, low state
β	0.996	Discount factor	λ_H	0	Skewness of transitory returns, high state
ϕ	2	Relative risk aversion	λ_L	-0.5	Skewness of transitory returns, low state
χ_1	9.4	Labor disutility	χ_1	0.01	Non-interest expense, linear
χ_2	1	Frisch elasticity	χ_2	1.45	Non-interest expense, quadratic
ν_1	0.52	Deposits in utility			
ν_2	0.1	Elasticity of deposit supply			
θ_b	2.2	CES markdown			

the standard deviation of the quarter-on-quarter RoL growth for the sample of commercial banks and over the usual time period.

Counter-cyclicality of loan return risk is shaped by two sets of parameters. First, the mean $\mu_{e,t}$ of the distribution of transitory risk takes the values of 0 (μ_h) in the high and -0.02 (μ_l) in the low aggregate states, respectively. This differential corresponds to the difference in mean quarter-on-quarter RoL growth across U.S. expansions and recessions, as shown on Panel (a) of Figure 2. Second, pro-cyclical skewness of transitory return risk is achieved by setting the values of $\lambda_{\epsilon,t}$ to 0 and -0.5 in high and low aggregate states, respectively. The value of -0.5. roughly approximates the average difference in skewness of RoL growth across U.S. expansions and recessions, as presented on Panel (b) of Figure 2. In other words, we have normalized the high aggregate state (expansion) in the model to feature no deviations from normality. Recessions, on the other hand, feature an extended left tail of the $\epsilon_t(j)$ density.

Finally, the diad $\{\chi_1, \chi_2\}$ is chosen in order to target two moments in the data. First, we calibrate χ_1 by targeting the aggregate book leverage ratio of banks, defined as the ratio of assets over equity. In the data, importantly, we define the numerator (assets) with total loans only. Under this definition, we find that the average leverage ratio is about 6.5 across banks and time. Leverage, when defined as total assets over equity, is considerably higher at around 11. In our model, the correct definition of leverage is loans over equity because there is only one risky asset in the economy and this asset most closely corresponds to a commercial loan to non-financial firms. Second, χ_2 is calibrated in order to match the volatility of aggregate consumption relative to that of aggregate output, which is about 0.62 in U.S. data over the usual time frame.

Summary Before proceeding with the presentation of main results, we briefly summarize the success and shortcomings of our calibration procedure. First, the model is designed to be flexible enough such we are able to hit our targets with relative ease. Second, many untargeted business cycle moments - such as the high correlation of output, consumption, and hours - are very much in line with the data. Third, bank size concentration in the stationary steady state, while considerably high, is still roughly 25% too low. This point has been discussed previously and potential remedies and extensions have been offered. Third, μ_t^k and μ_t^b are counter-cyclical, as in the data, but correlation coefficients with output seem too high in the model. This is driven by the fact that there is only one aggregate shock in our model and only one absorbing bank-level characteristic: size. Introducing a simple fix such as markup and markdown (i.i.d) shocks can bring the values to the empirical counterparts exactly. Fourth, bank balance sheet quantities and leverage in the model are - correctly - pro-cyclical but for the same reason as with markups and markdowns they are too pro-cyclical. This point is well understood since at least Nuno and Thomas (2017): financial sector's cyclicality properties are complicated and generally depend on the sample and exact definition of the "bank". Recall that our model-consistent empirical sample includes only depository institutions. The aggregate financial sector - inclusive of broker-dealers and finance companies - is much more volatile. Overall, financial and business cycle

5 Model Properties

We now investigate whether our calibrated and solved model is able to replicate key properties of the banking data.

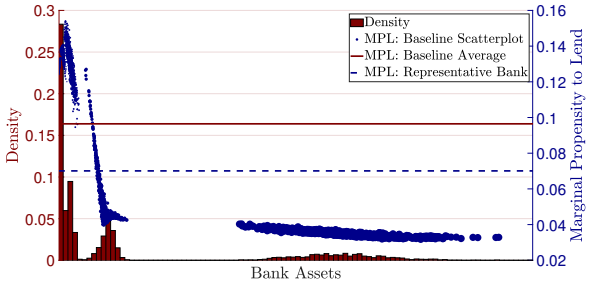
5.1 Bank Heterogeneity and Distributional Dynamics

We start with a central piece of our framework - the endogenous cross-sectional distribution of bank size. Figure 6 visualizes the distribution's two key aspects. Panel (a) shows the histogram of bank assets from the stationary equilibrium. The histogram reveals something akin to persistent multi-modality, which represents the behavior of permanent inequality types $\kappa(j)$. There is a small fraction of consistently large intermediaries, followed by medium- and many small-sized banks.

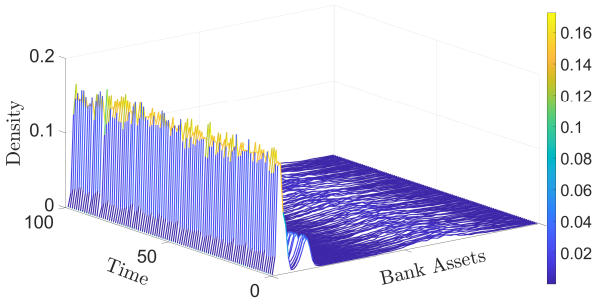
The same panel also shows heterogeneity in the marginal propensity to lend (MPL). We define the MPL as the change in bank-level lending $k(j)$ in response to a marginal change in net worth $n(j)$. From the Figure, we see that the MPL is noticeably decreasing in size, suggesting that small banks have a higher lending elasticity of net worth shocks. This property is consistent with the classic evidence on the heterogeneous effects of the bank lending channel (Kashyap and Stein, 1995, 2000). The Panel also reports

Figure 6: Bank Heterogeneity and Dynamics

(a) Marginal Propensity to Lend



(b) Distribution of Assets over Time



Notes:

unweighted-average MPL values for the baseline economy and for the representative-bank counterfactual. Note that the two must not necessarily equalize, since the average of the cross section does not need to equate the behavior of the representative agent model. In fact, this is precisely what occurs as the average MPL in the Bewley Banks economy is 9.7% while for the representative-bank benchmark it is roughly 7%. While the largest banks in our economy have low MPL, their share is not sufficiently high so that to counteract the larger mass of smaller, high-MPL banks that heterogeneity introduces. In anticipation of the next section, the MPL heterogeneity channel is the reason why our business cycle fluctuations - and particularly the response of total bank lending - get amplified in our economy[11].

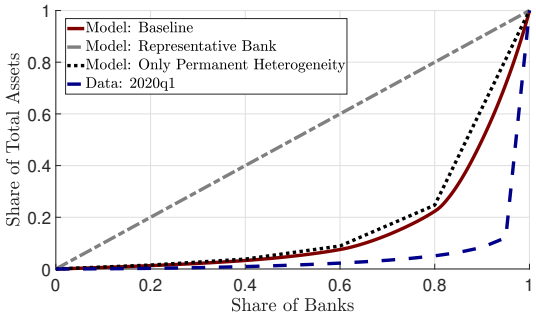
Panel (b) of Figure 6 presents a "waterfall" representation of how the full distribution of bank assets in our baseline economy evolves over time. The cross section is fully dynamic, and there is within-type cross-bank heterogeneity that is driven by transitory risk $\xi_t(j)$. Individual banks retain the average profitability of their business model over the long run but still move around the state space in the short run due to uninsured idiosyncratic shocks.

Figure 7 visualizes the model's ability to generate a realistic degree of size concentration. Panels (a) and (b) plot the model-implied Lorenz curves for $\int k(j)$ and $\int b(j)$, respectively, alongside the empirical counterparts. Recall that these object exist in equilibrium due to scale variance, which is in turn achieved through non-interest cost convexity. Any departure from the perfect equality counterfactual (the 45-degree line) suggests that either the loan or deposit market features an unequal degree of concentration. The baseline

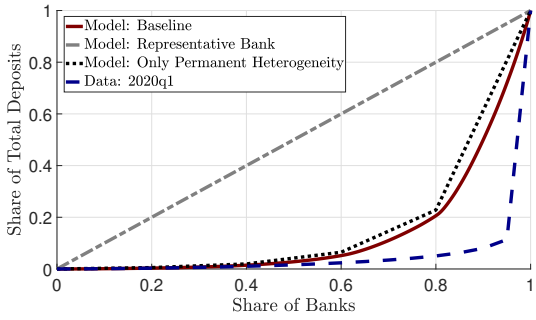
[11]Conceptually, our MPL object is related to several related constructs in the literature. For example, the Marginal Propensity to take on Risk (MPR) in an environment where heterogeneous households choose their portfolio risk exposure (Kekre and Lenel, 2022) and the "marginal propensity to invest" in models with heterogeneous firms and financial frictions (Ottonello and Winberry, 2020).

Figure 7: Bank Size Concentration

(a) Asset Market Concentration



(b) Deposit Market Concentration



Notes:

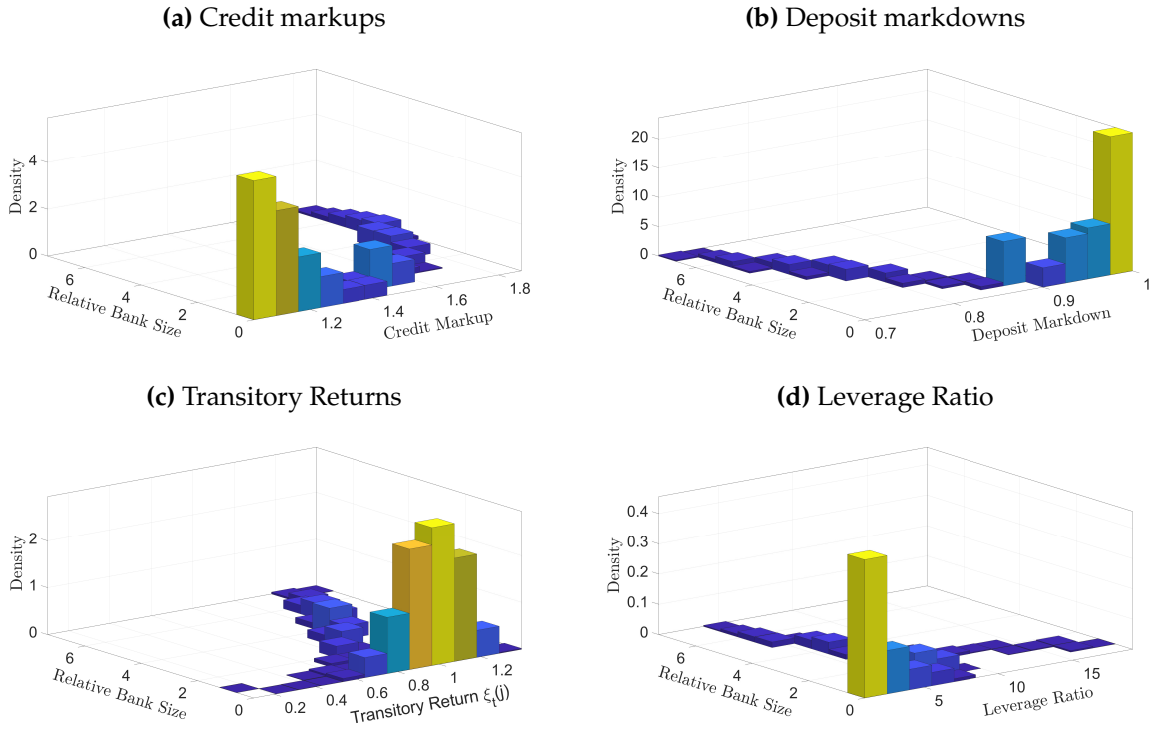
economy is considerably concentrated with Gini coefficients of 0.70 and 0.73 for assets and deposits, respectively. The representative-bank counterfactuals feature Gini coefficients of exactly 0, by construction. An important special case that is also shown on the same plots is the economy with only permament heterogeneity $\kappa(j)$, i.e., when the volatility of transitory risk σ_ϵ is set to zero. The Lorenz curves for the permanent-only cases are qualitatively non-distinguishable from the baseline. In other words, transitory risk does little to generate equilibrium size concentration while permanent inequality appears to be very important.

In the data, the Gini coefficient was roughly 0.94 in 2020:1. Despite a considerable, order-of-magnitude improvement over the representative-bank benchmark, our baseline economy still cannot account for the extreme levels of concentration in the U.S. banking sector. For example, it's difficult to engineer a situation where the top quintile of banks controls 95% of assets, as it is typically the case in the U.S., even when $\kappa(j)$ follows Zipf's law. In our baseline economy, the corresponding value is 78%. A relevant feature of the data that is missing in our framework is the mergers and acquisitions market, which has historically accounted for a non-trivial share of bank exits. Endogenous horizontal integration would allow large, high-profitability types to acquire franchises of small, low-type competitors. Introducing an M&A market could bring equilibrium concentration even closer to the data but is beyond the scope of this paper.

5.2 Cross-Sectional Correlations

Bank size is an essential characteristic in the model because it is also a relevant state. Every other variable can be computed if the distribution of net worth is pinned down. We are now interested in seeing whether the model predicts correct cross-sectional relationships between bank size and other key objects. Figure 8 presents a series of bivariate histograms

Figure 8: Cross-Sectional Correlations



Notes:

with relative banks' size $\left(\frac{k(j)}{K}\right)$ on the y-axis, a corresponding variable of interest on the x-axis, and density on the vertical z-axis. Panel (a) depicts the plot for credit markups $\mu^k(j)$. The relationship is increasing and concave, which is delivered by the choice of $\nu_1 < 0$ or the "keeping up with the Joneses" specification. The smallest banks in our model have essentially no market power, as evidenced by their choice of $\mu^k(j)$ which is close to unity. The largest banks, however, enjoy a lot of credit market power and can set gross markups in excess of 1.8 over their marginal costs. Panel (b) of the Figure shows a similar plot for deposit markdowns $\mu^b(j)$. Markdowns fall with size in an almost linear fashion, as they do also in the data. This relationship is guaranteed by $\theta_b > 0$. Whenever the household derives utility from liquidity holdings and as long as deposit franchises are imperfect substitutes, the larger bank will always hold more deposit market power because the household has a preference to save in larger banks. Panel (c) plots the realization of transitory risk draws $\xi(j)$ which is notably left-skewed as can be seen from the x-axis. Its correlation with relative size is strongly positive, intuitively suggesting that more profitable opportunities allow banks to accumulate more net worth. A significant mass of banks that are seemingly at the boundary of low size are of the low permanent types. For them, even abnormally positive transitory draws are not sufficient to compensate for a permanently low value of $\kappa(j)$.

Panel (d) of Figure 8 shows how the distribution of book leverage is shaped by size. The plot reveals an interesting non-linearity in the relationship. Starting from low levels of relative assets, book leverage marginally increases with size, as studies by Adrian and Shin (2010) and others have pointed out. At the extreme, however, the largest banks have only moderate leverage ratios of 5-7. These institutions have outgrown all constraints - either in the form of the hard leverage constraint or the pressure of convex asset adjustment costs. The largest banks in our economy are safe from the financial stability point of view. At the same time, there is also a non-trivial share of small banks with very high leverage ratios. These franchises are not profitable, have low levels of net worth, and cannot outgrow the cost constraint. Importantly, a relevant notion of "risk" in our economy would correspond to insolvency risk, which is cancelled out by the deposit insurance scheme but could still matter ex-post if, for example, default was costly. Since small banks are closer to fundamental insolvency, we do obtain concentration of leverage and fundamental risk (Coimbra and Rey, 2023). This concentration, however, is in the left tail of the size distribution. Importantly, we abstract from bank runs (Gertler et al., 2016; Amador and Bianchi, 2021) or the presence of a "too big to fail" externality (Philippon and Wang, 2022). Both frictions, especially the latter by definition, would disproportionally impact large banks, thus skewing risk towards the right tail.

6 Aggregate Fluctuations

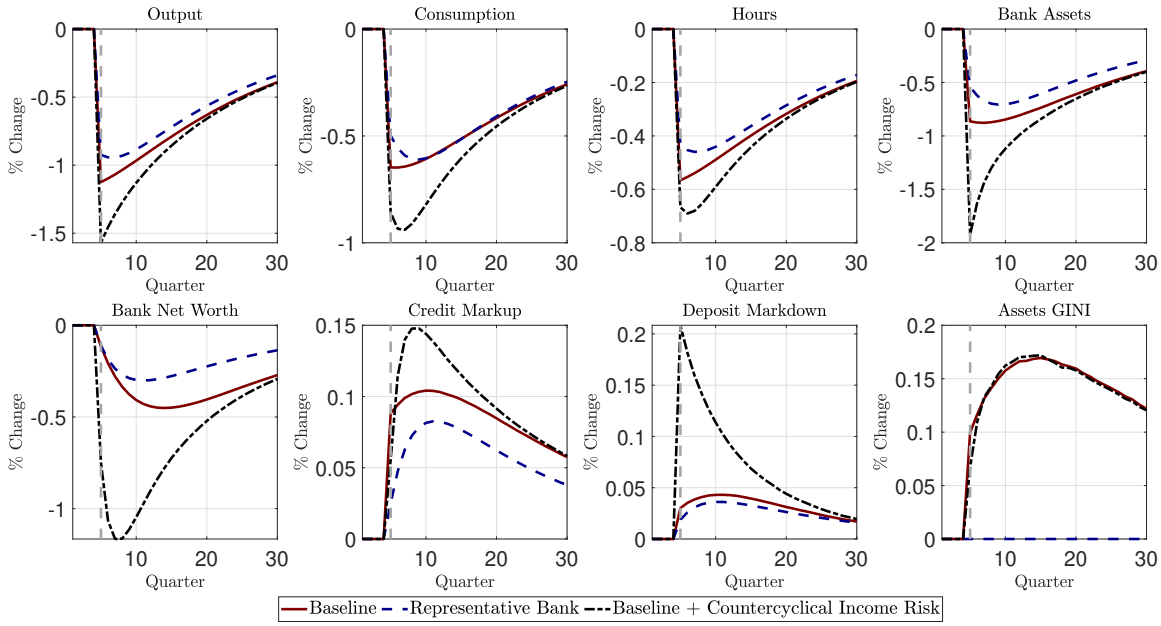
In this section we characterize model responses to aggregate and idiosyncratic shocks. We begin by studying model behavior in response to aggregate TFP shocks. We proceed by investing the mechanism and isolating the contributions of different frictions. Finally, we characterize model behavior in response to granular shocks, i.e., idiosyncratic disturbances that hit only a specified subset of banks.

To obtain the plotted impulse responses with respect to an aggregate shock, we perform the following computational steps. First, we run a simulation based on an already-solved economy with **I** banks for **T** periods using only idiosyncratic shocks up until quarter T^* - represented on the figures by the vertical line. At time T^* there is a quarterly innovation to A_t that mean-reverts back to unity over time. Second, we run a second simulation which is identical to the first except that there is not aggregate shock hitting at T^*. Third, to obtain impulse responses we subtract the path of aggregates in the second simulation from the path of aggregates in the first simulation.

6.1 Counter-Cyclical Idiosyncratic Risk

Figure 9 displays the effects of a one-standard deviation negative shock to aggregate TFP. We compare three model specifications. First, the baseline case of heterogeneous banks

Figure 9: The Role of Counter-Cyclical Risk



Notes: impulse responses to an aggregate TFP shock. Baseline model with a-cyclical income risk (solid) vs. representative bank (dashed) vs. baseline model with counter-cyclical income risk (dotted)

with *a-cyclical* idiosyncratic risk (solid line). Second, the case of heterogeneous banks with *counter-cyclical* idiosyncratic risk (dotted line). Third, the limit case of a representative bank (dashed line) which coincides with the GK economy.

In the baseline economy, a negative TFP shock generates a severe financial tightening driven by a decline in the aggregate return on investment: the size of the banking sector shrinks as both assets and net worth fall. The bank lending channel transmits onto non-financial firms which, receiving less funding from the banks, produce less capital. As a result, an economic recession ensues: total output, consumption, and hours all decline.

In line with our empirical evidence, both credit markups and deposit markdowns increase. To gain intuition, it is useful to first focus on the banks' liability side. Since the risk-free rate is rising (and consumption falling), banks face an increase in the household's opportunity cost of holding deposits relative to mutual funds. To prevent deposit withdrawals, banks raise their deposit rates more than proportionally relative to the risk-free rate, leading to a counter-cyclical rise in the markdown, in line with what we observed in the data. This is consistent with the household's marginal liquidity preference for deposits falling in a bad aggregate state. The rise in the deposit rate leads to an increase in banks' marginal cost (although less than proportional, since the non-interest component of the marginal cost is contracting as banks are shrinking in size). In turn, on the asset side, banks shield their profits by raising the credit rate more than proportionally relative to

their marginal cost, leading to a counter-cyclical rise in the credit markup. Interestingly, and precisely because the deposit rate is an important component of banks' marginal cost, there is a strict interaction between market power on the liability side and market power on the asset side of banks' balance sheets. Furthermore, notice that the banking sector is becoming more concentrated, as seen from the rising GINI coefficient.

Relative to the behavior of the representative-bank case, the baseline economy exhibits noticeable amplification of both financial and macroeconomic aggregates, particularly bank assets and total output. The intuition for this result is best understood by recalling heterogeneity in the marginal propensity to lend (MPL). As Figure 6 had documented, the average MPL in the baseline economy is higher than the MPL of the representative (GK) intermediary. As such, this means that the baseline economy is more susceptible to aggregate shocks.

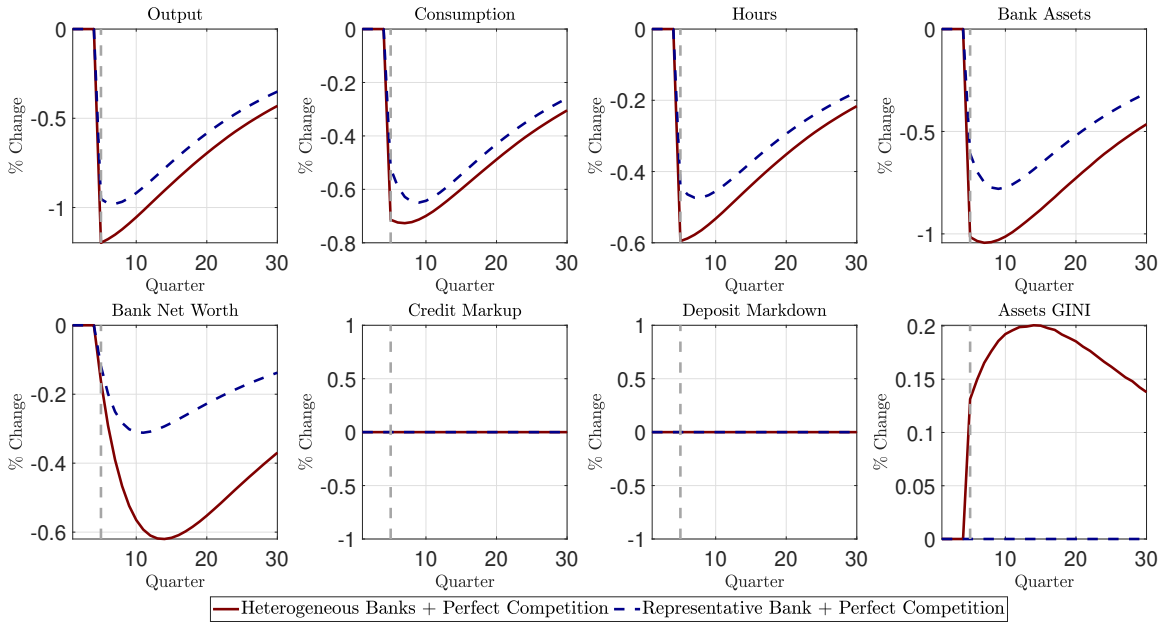
An important result from Figure 9 concerns the economy with counter-cyclical income risk. Recall that in this case, as soon as the negative aggregate shock hits the economy, both $\mu_{\epsilon,t}$ and $\lambda_{\epsilon,t}$ switch permanently to their corresponding low-state values such that banks now draw from a more left-skewed density of $\xi(j)$. Counter-cyclical risk generates significant amplification of aggregate fluctuations, *on top* of whatever amplification heterogeneity adds to the representative-bank case. The response of output is up to fifty percent larger in the Bewley Banks economy with counter-cyclical risk relative to the baseline with a-cyclical risk. The financial crisis is also much more severe and pronounced as bank assets and net worth collapse by 100% more. Counter-cyclicality of bank income risk - a robust feature of the micro data - appears to be an important source of financial and business cycle amplification.

6.2 Heterogeneity

Does bank heterogeneity *per se* matter for aggregate fluctuations? We now perform a test that isolates the role of heterogeneity by shutting down bank market power, which may have conflated our conclusions from Figure 9. Figure 10 displays the effects of an aggregate negative TFP shock in two counter-factual cases: a heterogeneous bank perfect competition economy (HBPC, solid line) and an economy with a representative perfectly competitive bank (RBPC, dashed line). Thus, the latter case de facto corresponds to the GK model. Recall that obtaining the HBPC economy entails (a) shutting down deposit liquidity preferences by setting $\nu_1 = 0$ and (b) using the CES aggregator in (4) by choosing $\gamma_1 = 0$.

Figure 10 confirms that bank heterogeneity *per se* leads to amplification of aggregate shocks: output contracts by up to 25% per cent more in the case of bank heterogeneity relative to the representative-bank limit. As already suggested earlier, the intuition for why heterogeneity leads to amplification lies in the presence of a large share of small

Figure 10: The Role of Bank Heterogeneity



Notes: impulse responses to a negative aggregate TFP shock. Heterogeneous banks and perfect competition (solid) vs. representative bank and perfect competition (dashed).

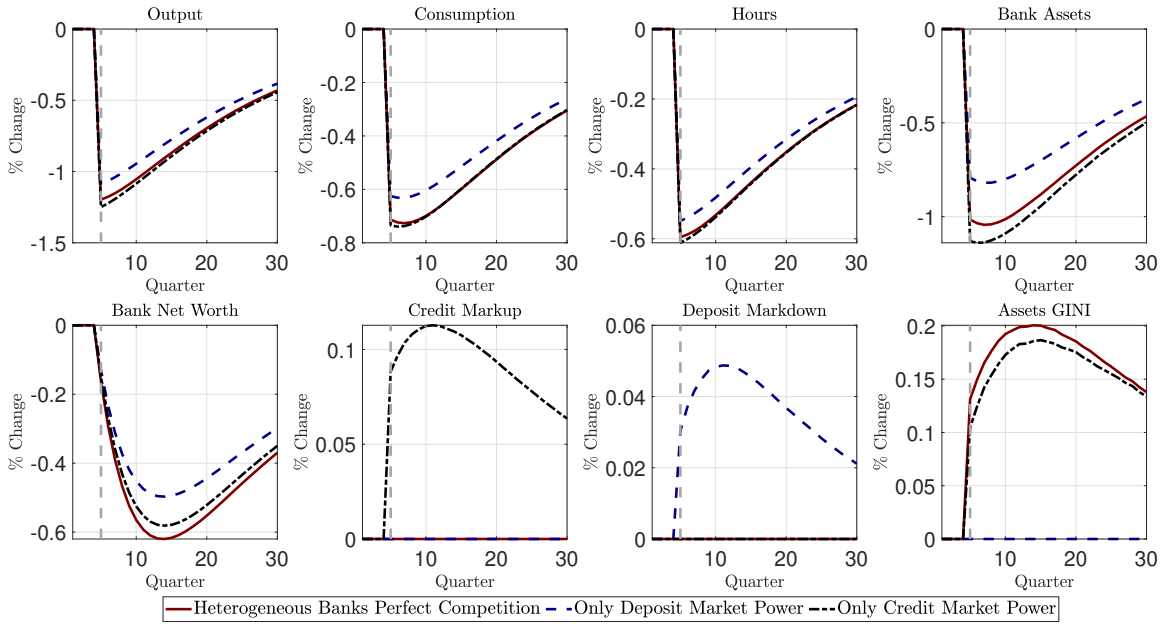
banks with a large MPL, leading to a greater average ex-ante economywide MPL and a substantial ex-post decline in net worth and bank assets[12].

6.3 Market Power

In Figure 11 we inspect the role of two-sided bank market power. We compare impulse responses to an aggregate shock in three alternative cases: (i) heterogeneous banks with perfectly competitive credit and deposit markets (solid line); (ii) heterogeneous banks with market power only on the credit side (dotted line); (iii) and heterogeneous banks with market power only on the deposit side (dashed line). We find that credit market power amplifies (although marginally) both real and financial aggregate fluctuations. The intuition for this finding stems from credit markups being counter-cyclical. As discussed earlier in the paper, in recessions banks raise credit rates more than proportionally relative to marginal costs in order to compensate for falling franchise values with higher short-run profits. The larger the credit rate increase following a negative aggregate shock, the larger the decline in asset quantities and net worth. In turn, output and consumption all fall to

[12]This intuition is analogous to the logic of a large class of models with heterogeneous households, such as in the influential two-agent and heterogeneous-agent New Keynesian (TA/HANK) literature (Galí et al., 2007; Bilbiie, 2008; McKay and Reis, 2016; Kaplan et al., 2018; Hagedorn et al., 2019).

Figure 11: The Role of Market Power



Notes: impulse responses to a negative aggregate TFP shock. Perfect competition (solid) vs. deposit market power only (dashed) vs. credit market power only (dotted)

a larger extent relatively to the case of perfect credit markets.

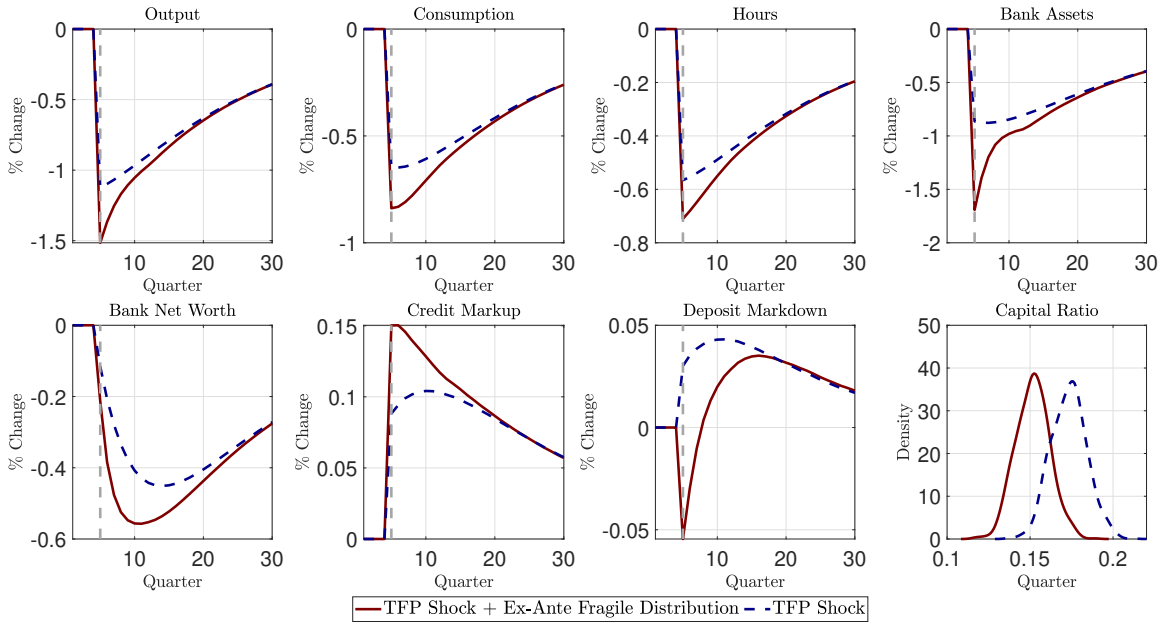
Conversely, deposit market power dampens aggregate fluctuations. In a recession, since the risk-free rate is rising, the households' opportunity cost of saving via deposits rises. Hence, households have the incentive to shift demand away from deposits towards mutual funds, whose return has increased. Banks try to preserve the demand for deposits by raising deposit rates more than proportionally relative to the risk-free rate. As a result depositors (households) are better able to smooth consumption, which leads to a more muted contraction in output relative to the case of perfect deposit markets. Quantitatively, the dampening effect of deposit market power is substantive, suggesting that the deposit franchise is an important hedging resource for banks to navigate the cycle.

6.4 Fragile Bank Distribution

In our environment, the endogenous distribution of bank net worth is an important, relevant state variable. As such, all aggregate responses to exogenous shocks depend explicitly on its shape and condition. In this vein, business cycles can be a function of the underlying degree of financial fragility of the banking cross section. In other words, our model should be able to generate distributional aggregate state-dependency.

We operationalize this idea by comparing aggregate dynamics in the baseline case with what we label as the "fragile" economy. The fragile economy is such that, in period $T^* - 1$,

Figure 12: Fragile Banking Distribution

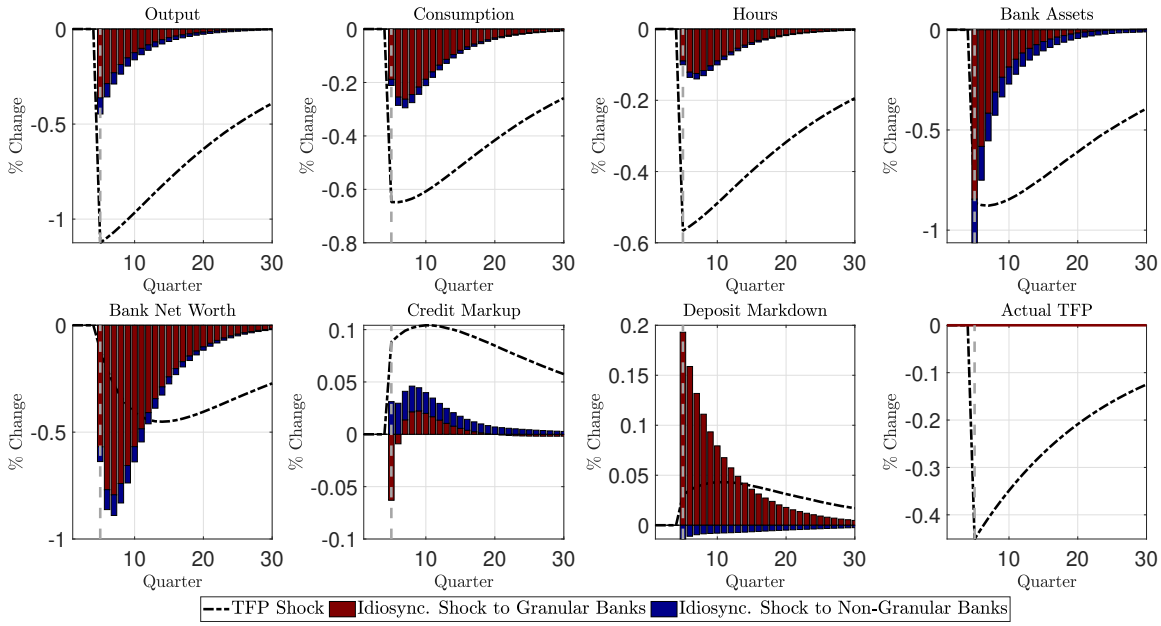


Notes: impulse responses to a negative aggregate TFP shock. Fragile (dashed) vs. non-fragile banking distribution (solid)

a negative financial shock has caused a leftward shift in the distribution of bank capital ratios - defined as book net worth over book assets. This distribution is characterized by a lower mean and a higher right-skewness. We obtain impulse responses for the fragile economy in a similar way as before. We run two model simulations, both using only idiosyncratic shocks and with the fragility-inducing financial shock occurring at $T^* - 1$. We assume that this shock lasts for 2 quarters. We further assume that the negative TFP shock takes place at T^* in the first simulation but not in the second. Taking the difference in responses across the two simulations gives us the desired result.

Figure 12 plots impulse responses to an aggregate negative TFP shock under two different scenarios. In this experiment, a negative aggregate shock that occurs once the banking sector is already fragile generates a significantly more severe financial and real-economy contractions. The excess contraction scales with the duration and severity of the prior financial shock. The mechanism for this outcome relies on the MPL heterogeneity logic: the fragile economy features a higher starting average MPL because a greater number of banks are close to very low or zero net worth. As a result, any subsequent negative aggregate shock has larger effects on the economy.

Figure 13: The Role of Granularity



Notes: impulse responses to granular return shocks (red histogram) vs. idiosyncratic return shocks to all banks (blue histogram) vs. aggregate TFP shock (dashed line)

6.5 Granular Banks

We have seen previously that, conditional on counter-cyclical mean and pro-cyclical skewness, idiosyncratic disturbances significantly amplify aggregate fluctuations. But can idiosyncratic shocks only to *granular* banks, i.e., banks in the top quintile of the size distribution, generate a realistic-looking business cycle? In other words, we are now testing explicity whether our model can delivery granular bank dynamics.

In Figure 13, we show the results of an exercise that involves comparing impulse responses across three sets of specifications. First, we simulate an economy with only idiosyncratic shocks. In period T^*, only banks in the largest quintile of the distribution of assets experience a drop in the mean and in the skewness of the idiosyncratic shock density. We label this as the "granular shock" and represent it with red histograms on the Figure. Nothing changes for the remaining four quintiles and aggregate TFP remains constant. In an alternative simulation, the granular shock does not takes place. Taking differences across simulations delivers the first set of impulse response functions (IRFs). Second, we simulate an economy with only idiosyncratic shocks and in period T^* *all* banks experience a drop in the mean and in the skewness of the idiosyncratic shock density, all the while aggregate TFP remains constant. In a second simulation, the countercyclical income risk shock does not take place and taking differences across simulations gives the second set of IRFs. This case is represented with blue histograms on the Figure. Finally,

the third case is the usual simulation of a one-standard deviation negative TFP shock that occurs in T^*, as discussed before.

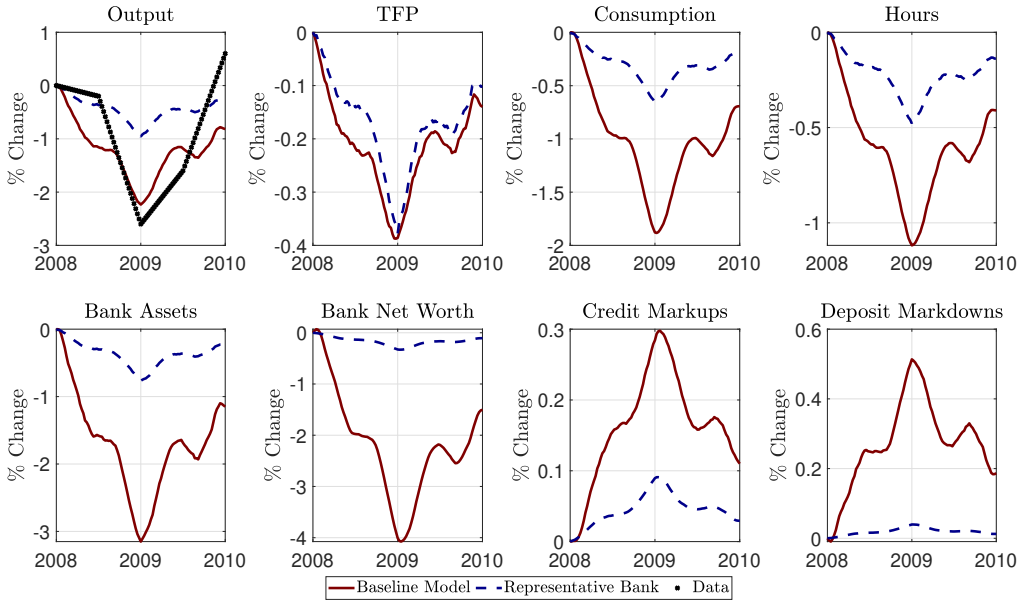
Two observations are in order. First, granular shocks can account for almost all of the variation in real and financial variables induced by idiosyncratic shocks. Hence, the top quintile of the size distribution acts as a "sufficient statistic" for the impact of bank-level idiosyncratic shocks on the business cycle. Second, granular banking shocks cause a contraction in output that is about 40% of that generated by a negative aggregate TFP shock. In addition, granular shocks account for a significantly larger share of the response of financial variables - assets and net worth - relative to the case of an aggregate shock. The bank size distribution in our economy is very concentrated - as it is in the data. A small share of intermediaries controls a substantive share of both loans and deposits. Idiosyncratic disturbances to those largest banks can by themselves generate realistic financial and business cycles. As a result, our model endogenously generates what regulators in practice call "systemically important banks" or SIBs.

7 Banking and Economic Crises

In this section we use our model to characterize banking and economic crises. We employ an event study approach. We simulate any given economy for 5,000 periods and define an event as a decline in output of 2.5 standard deviations below the average. This is chosen because the contraction in output during the Great Recession was roughly a 2.5 standard-deviation shock. We store the path of every relevant variable on the interval of some periods before to some periods after the event occurs. Then, we take the averages across events.

In Figure 14 we compare the baseline model with heterogeneous banks and counter-cyclical idiosyncratic risk (solid line) with a model characterized by only a representative bank (dashed line). As it is clear the economy with heterogeneous banks is able to track the behavior of output observed during the Great Recession (dotted line) relatively well, whereas the economy with a representative bank is able to account for less than half of the trough in output. In general, the model with heterogeneous banks generates much steeper contractions in banks' assets and net worth around a financial crisis episode relative to the model with only a representative bank. The reason for why financial crises are sudden and sharper in the economy with heterogeneous banks and idiosyncratic shocks is twofold. First, idiosyncratic risk is counter-cyclical. Second, the average MPL is larger than the MPL in the economy with only a representative bank.

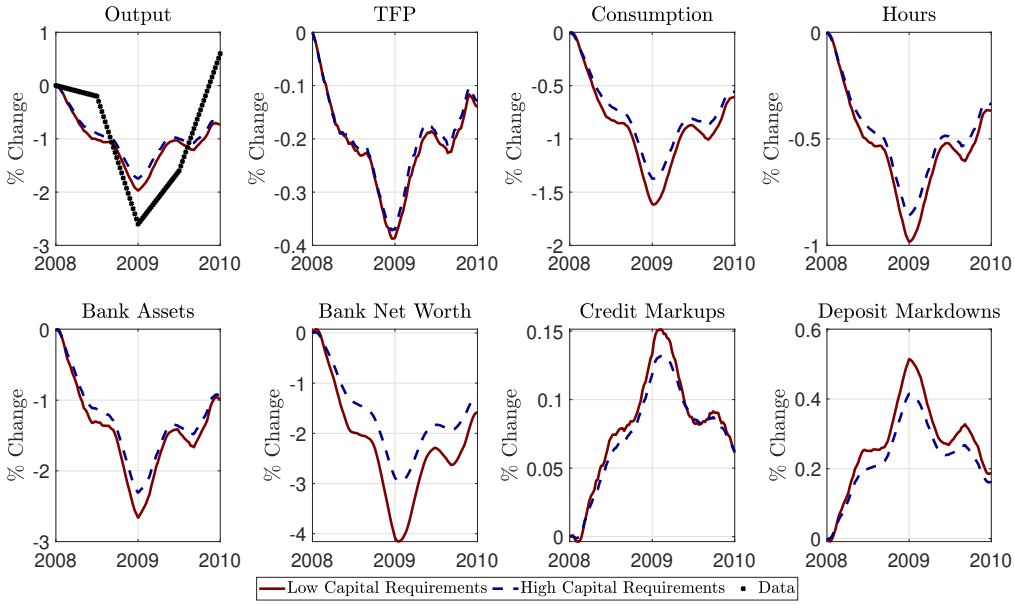
Figure 14: Banking and Economic Crises



Notes: banking crisis event under the baseline model (solid line) vs crisis event under a representative bank model (dashed line) vs data (dotted line).

Capital Requirements Finally, we turn to the role of macroprudential regulation, which has been the object of a large literature in macroeconomics and banking. We study whether capital requirements can mitigate the loss of output during crises. Figure **??** illustrates the results of this experiment. We present two alternative crisis scenarios. In one scenario (solid line) the banking sector is characterized by low capital requirements, whereas in the alternative scenario (dashed line) capital requirements are doubled. Doubling capital requirements improves financial stability and dampens economic contractions in typical model-simulated recessions. The reduction is, however, quantitatively small and is potentially not justified given that capital requirements generate lower levels of lending and production in the high-regulation steady state. **TM EXPAND THIS SECTION?**

Figure 15: Capital Requirements and Crises



Notes: banking crisis event with low capital requirements (solid line) vs banking crisis event with high capital requirements (dashed line) vs data (dotted line).

8 Conclusions

We have developed a new tractable, dynamic stochastic general equilibrium framework with two-sided 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, a feature that typically characterizes the models with a representative intermediary. Because the marginal value of net worth and the optimal leverage ratios are now both size-dependent, a time-varying distribution of bank characteristics emerges in equilibrium. With aggregate uncertainty, the distribution of bank net worth becomes a time-varying endogenous state variable.

Our framework matches a series of stylized facts of the banking data: (i) heterogeneity and right-skewness in the distribution of banks' assets and deposits; (ii) a power-law distribution of bank size; (iii) a significant degree of granularity; (iv) time-varying two-sided market power, and (v) counter-cyclical idiosyncratic rate of return risk, resulting from pro-cyclical skewness. Our Bewley Banks model is consistent with the empirical properties (i)-(v). Relative to a model with a representative bank, the model delivers significant amplification of real and financial variables in response to aggregate shocks. This is due to

two main forces simultaneously at play: pro-cyclical skewness of idiosyncratic risk, and an average marginal propensity to lend which is larger than the one in the corresponding representative bank economy.

Our Bewley Banks framework is tractable and portable. We envision at least three extensions for future research. First, introducing nominal rigidities to study how bank heterogeneity affects the transmission mechanism of monetary policy. Second, using our Bewley Banks environment to study unconventional credit policies such as bank-level equity injections, possibly when the economy has hit the effective lower bound on interest rates. Third, relaxing the closed economy assumption to characterize the different behavior of domestic vs. global banks.[13]

References

Abadi, J., M. Brunnermeier, and Y. Koby, "The Reversal Interest Rate," *Working Paper*, 2022.

Adrian, T. and H. S. Shin, "Liquidity and Leverage," *Journal of Financial Intermediation*, 2010, *19(3)*, 418–437.

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

Allen, F. and D. Gale, "Optimal Financial Crises," *Journal of Finance*, 1998, *53(4)*.

_ **and** _ , "Financial Intermediaries and Markets," *Econometrica*, 2004, *72(4)*.

Amador, M. and J. Bianchi, "Bank Runs, Fragility, and Credit Easing," *NBER Working Paper*, 2021, *29397*.

Baltagi, B. and P. Wu, "Unequally Spaced Panel Data Regressions with AR(1) Disturbances," *Econometric Theory*, 1999, *15*.

Begenau, J. and T. Landvoigt, "Financial Regulation in a Quantitative Model of the Modern Banking System," *Review of Economic Studies*, 2021, *89*.

_ **, S. Bigio, J. Majerovitz, and M. Vieyra**, "A Q-Theory of Banks," *NBER Wokring Paper*, 2021, *27935*.

Bellifemine, M., R. Jamilov, and T. Monacelli, "HBANK: Monetary Policy with Heterogeneous Banks," *CEPR Working Paper*, 2022, *17129*.

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

_ **and** _ , "Skewed Wealth Distributions: Theory and Empirics," *Journal of Economic Literature*, 2018, *56*.

Bernanke, B. and A. Blinder, "Credit, Money, and Aggregate Demand," *American Economic Review*, 1988, *78*.

_ **and M. Gertler**, "Inside the Black Box: The Credit Channel of Monetary Policy Transmission," *American Economic Review*, 1995, *9*.

[13] In Bellifemine et al. (2022) we study the monetary policy transmission mechanism in a Bewley Banks environment with nominal rigidities. In Cesa-Bianchi et al. (2023) we characterize empirically the heterogeneity between domestic and global banks, and study their behavior in an open-economy Bewley Banks model.

_ , _ , **and S. Gilchrist**, "The financial accelerator in a quantitative business cycle framework," *Handbook of Macroeconomics*, 1999, *1.*

Bewley, Truman, "The permanent income hypothesis: A theoretical formulation," *Journal of Economic Theory*, 1977, *16* (2), 252 – 292.

Bianchi, J. and S. Bigio, "Banks, Liquidity Management and Monetary Policy," *Econometrica*, 2022, *90.*

Bilbiie, F., "Limited asset markets participation, monetary policy and (inverted) aggregate demand logic," *Journal of Economic Theory*, 2008, *140* (1), 162–196.

Bloom, N., F. Guvenen, and S. Salgado, "Skewed Business Cycles," *NBER Working Paper*, 2019, *26565.*

Bocola, L., "The Pass-Through of Sovereign Risk," *Journal of Political Economy*, 2016, *124.*

_ **and G. Lorenzoni**, "Risk-Sharing Externalities," *Journal of Political Economy*, 2023, *131.*

Boissay, F., F. Collard, and F. Smets, "Booms and Banking Crises," *Journal of Political Economy*, 2016, *124(2).*

Boppart, T., P. Krusell, and K. Mitman, "Exploiting MIT shocks in heterogeneous-agent economies: the impulse response as a numerical derivative," *Journal of Economic Dynamics and Control*, 2018, *89.*

Boyd, J. and G. De Nicolo, "The Theory of Bank Risk Taking and Competition Revisited," *Journal of Finance*, 2005, *60(3).*

Broer, T., A. Kohlhas, K. Mitman, and K. Schlafmann, "Expectation and Wealth Heterogeneity in the Macroeconomy," *Working Paper*, 2022.

_ , _ , _ , **and** _ , "On the possibility of Krusell-Smith Equilibria," *Journal of Economic Dynamics and Controls*, 2022, *141.*

Brunnermeier, M. and L. Pedersen, "Market Liquidity and Funding Liquidity," *Refiew of Financial Studies*, 2009, *22*, 2201–2238.

_ **and Y. Sannikov**, "A Macroeconomic Model with a Financial Sector," *American Economic Review*, 2014, *104(2)*, 379–421.

Buch, C., D. Domeij, F. Guvenen, and R. Madera, "Skewed Idiosyncratic Income Risk over the Business Cycle: Sources and Insurance," *American Economic Journal: Macroeconomics*, 2022, *14(2).*

Camanho, N., H. Hau, and H. Rey, "Global portfolio rebalancing and exchange rates," *Review of Financial Studies*, 2022, *35.*

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

Chevalier, J. and D. Scharfstein, "Capital-Market Imperfections and Countercyclical Markups: Theory and Evidence," *American Economic Review*, 1996, *86(4).*

Christiano, Lawrence and Daisuke Ikeda, "Leverage Restrictions in a Business Cycle Model," *NBER Working Paper 18688*, 2013.

Coimbra, N. and H. Rey, "Financial Cycles with Heterogeneous Intermediaries," *The Review of Economic Studies*, 2023, *Forthcoming.*

Cooley, T. and V. Quadrini, "Financial Markets and Firm Dynamics," *American Economic Review*, 2001, *91.*

Corbae, D. and P. D'Erasmo, "Rising bank concentration," *Journal of Economic Dynamics and Control*, 2020, *115.*

_ **and** _ , "Capital Requirements in a Quantitative Model of Banking Industry Dynamics,"

Econometrica, 2021, *89(6).*

_ **and** _ , "Banking Industry Dynamics Across Time and Space," *Working Paper*, 2022.

Cuciniello, V. and F. Signoretti, "Large Banks, Loan Rate Markup, and Monetary Policy," *International Journal of Central Banking*, 2015, *11(3).*

Cúrdia, V. and M. Woodford, "Credit Spreads and Monetary Policy," *American Economic Review*, 2001, *91.*

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

_ **and P. Dybvig**, "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, 1983, *91(3).*

Drechsler, I., A. Savov, and P. Schnabl, "The deposits channel of monetary policy," *Quarterly Journal of Economics*, 2017, *132* (4), 1819–1876.

_ **,** _ **, and** _ , "Banking on Deposits: Maturity Transformation without Interest Rate Risk," *Journal of Finance*, 2021, *76.*

Egan, M., A. Hortacsu, and G. Matvos, "Deposit Competition and Financial Fragility: Evidence from the US Banking Sector," *American Economic Review*, 2017, *107(1).*

Elenev, V., R. Lanvoigt, and S. Van Nieuwerburgh, "A Macroeconomic Model With Financially Constrained Producers and Intermediaries," *Econometrica*, 2021, *89.*

Farhi, E. and J. Tirole, "Shadow Banking and the Four Pillars of Traditional Financial Intermediation," *Review of Economic Studies*, 2020, *88(6).*

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

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

Galaasen, S., R. Jamilov, R. Juelsrud, and H. Rey, "Granular Credit Risk," *Working Paper*, 2020.

Gali, J., "Keeping up with the Joneses: Consumption Externalities, Portfolio Choice, and Asset Prices," *Journal of Money, Credit, and Banking*, 1994, *26(1).*

_ , "Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework and Its Applications," *Princeton University Press*, 2008.

Galí, J., D. López-Salido, and J. Vallés, "Understanding the Effects of Government Spending on Consumption," *Journal of the European Economic Association*, 2007, *5(1).*

Gaubert, Cecile and O. Itskhoki, "Granular Comparative Advantage," *Journal of Political Economy*, 2021, *129.*

Gerali, A., S. Neri, L. Sessa, and F. Signoretti, "Credit and Banking in a DSGE Model of the Euro Area," *Journal of Money, Credit, and Banking*, 2010, *42(s1).*

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

_ **and P. Karadi**, "A Model of Unconventional Monetary Policy," *Journal of Monetary Economics*, 2011, *58(1)*, 17–34.

_ **, N. Kiyotaki, and A. Prestipino**, "Wholesale Banking and Bank Runs in Macroeconomic Modelling of Financial Crises," *Handbook of Macroeconomics*, 2016, *2.*

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

Goldstein, Itay, Alexandr Kopytov, Lin Shen, and Haotian Xiang, "Synchronicity and Fragility," *Working Paper*, 2022.

Grassi, B., M. De Ridder, and G. Morzenti, "The Hitchhiker's Guide to Markup Estima-

tio," *CEPR Discussion Paper*, 2022, *17532.*

Greenwood, J., Z. Hercowitz, and G. Huffman, "Investment, Capacity Utilization, and the Real Business Cycle," *American Economic Review*, 1988, *78(3).*

Guvenen, F., B. Kuruscu, S. Ocampo, and D. Chen, "Use it or Lose it: Efficiency and Redistributional Effects of Wealth Taxation," *Quarterly Journal of Economics*, 2023, *Forthcoming*.

_ , S. Ozkan, and J. Song, "The Nature of Countercyclical Income Risk," *Journal of Political Economy*, 2014, *122(3)*, 621–660.

Hagedorn, M., I. Manovskii, and K. Mitman, "The Fiscal Multiplier," *NBER Working Paper 25571*, 2019.

Hansen, B., "Autoregressive Conditional Density Estimation," *International Economic Review*, 1994, *35*, 705–730.

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

Heider, F., F. Saidi, and G. Schepens, "Life below Zero: Bank Lending under Negative Policy Rates," *The Review of Financial Studies*, 2019, *32.*

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

Holstrom, B. and J. Tirole, "Financial Intermediation, Loanable Funds, and the Real Sector," *Quarterly Journal of Economics*, 1997, *112(3)*, 663–691.

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

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

Jamilov, R., "A Macroeconomic Model with Heterogeneous Banks," *Working Paper*, 2020.

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

Kaplan, Greg, Benjamin Moll, and Giovanni L. Violante, "Monetary Policy According to HANK," *American Economic Review*, 2018, *108* (3).

Kashyap, A. and J. Stein, ""The impact of monetary policy on bank balance sheets," *Carnegie-Rochester Conference Series on Public Policy*, 1995, *42.*

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

Kekre, R. and M. Lenel, "Monetary Policy, Redistribution, and Risk Premia," *Econometrica*, 2022, *90.*

Kiyotaki, N. and J. Moore, "Credit Cycles," *Journal of Political Economy*, 1997, *105(2)*, 211–248.

Krusell, P. and A. Smith, "Income and Wealth Heterogeneity, Portfolio Choice, and Equilibrium Asset Returns," *Macroeconomic Dynamics*, 1996, *1*, 387–422.

_ and _ , "Income and Wealth Heterogeneity in the Macroeconomy," *Journal of Political Economy*, 1998, *106*, 867–896.

Kurlat, P., "Deposit spreads and the welfare cost of inflation," *Journal of Monetary Economics*, 2019, *106.*

Lee, S., R. Luetticke, and M. Ravn, "Financial Frictions: Macro vs Micro Volatility," *CEPR*

DP, 2020, *15133.*

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

Martinez-Miera, David and Rafael Repullo, "Does Competition Reduce the Risk of Bank Failure?," *The Review of Financial Studies*, 2010, *23* (10).

McKay, Alisdair and Ricardo Reis, "The Role of Automatic Stabilizers in the U.S. Business Cycle," *Econometrica*, 2016, *84* (1), 141–194.

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

Mendoza, Enrique G., "Sudden Stops, Financial Crises, and Leverage," *American Economic Review*, 2010, *100* (5).

Nekarda, C. and V. Ramey, "The Cyclical Behavior of the Price-Cost Markup," *Journal of Money, Credit and Banking*, 2021, *52(2).*

Nuno, G. and C. Thomas, "Bank Leverage Cycles," *American Economic Journal: Macroeconomics*, 2017, *9(2).*

_ , J. Fernandez-Villaverde, and S. Hurtado, ""Financial Frictions and the Wealth Distribution," *Econometrica*, 2023, *Forthcoming.*

Ottonello, P. and T. Winberry, "Financial Heterogeneity and the Investment Channel of Monetary Policy," *Econometrica*, 2020, *88(6).*

Philippon, T. and O. Wang, "Let the Worst One Fail: A Credible Solution to the Too-Big-To-Fail Conundrum," *Quarterly Journal of Economics*, 2022, *Forthcoming.*

Pollak, R., "Additive Utility Functions and Linear Engel Curves," *Review of Economic Studies*, 1971.

Polo, Alberto, "Imperfect pass-through to deposit rates and monetary policy transmission," *Bank of England staff working papers*, July 2021, *No. 933.*

Ravn, M., S. Schmitt-Grohe, and M. Uribe, "Deep Habits," *The Review of Economic Studies*, 2006, *73(1).*

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

Sidrauski, M., "Inflation and Economic Growth," *Journal of Political Economy*, 1967, *75.*

Smets, F. and R. Wouters, "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, 2007, *97(3).*

Tella, S. Di and P. Kurlat, "Why Are Banks Exposed to Monetary Policy?," *American Economic Journal: Macroeconomics*, 2021, *13.*

Walsh, C., "Monetary Theory and Policy," *The MIT Press*, 2010.

Wang, O., "Banks, Low Interest Rates, and Monetary Policy Transmission," *Working Paper*, 2022.

Wang, Yifei, Toni M Whited, Yufeng Wu, and Kairong Xiao, "Bank Market Power and Monetary Policy Transmission: Evidence from a Structural Estimation," *Journal of Finance*, 2022, *77(4).*

Whited, Toni M., Yufeng Wu, and Kairong Xiao, "Low interest rates and risk incentives for banks with market power," *Journal of Monetary Economics*, 2021, *121*, 155–174.

Online Appendix for "Bewley Banks"

Rustam Jamilov Tommaso Monacelli

April 10, 2023

Contents

A Empirical Appendix

A.1 Data Details

This section provides details on our data work. Table A1 summarizes all series that are used throughout the paper. The main source of data for us is the Consolidated Report of Condition and Income, known as the Call Reports. This dataset covers all U.S. banks that are regulated by the Federal Deposit Insurance Corporation (FDIC). We focus on commercial banks, a list that includes depository trust companies, credit card companies with commercial bank charters, private banks, development banks, and limited charter banks. The sample runs over the 1984:1-2020:1 period. The level of aggregation is on an individual bank level, identified with the Federal Reserve identifier (RSSID). Throughout, we restrict the sample to observations with non-negative equity (RCFD3210). We have identified bank exits that are due to mergers or acquisitions using the Call Reports' Transformation Table and control for them by discarding observations when they occur.

Our measure of U.S. GDP growth, which is shown on Figure 1, is real Gross Domestic Product obtained from the St. Louis Federal Reserve (FRED database). The series has been logged and filtered with the Hodrick-Prescott filter under the usual smoothing parameter 1,600. Our measure of total bank assets is the variable RCFD2170, deflated with the CPI index. Estimated loan markups and deposit markdowns, plotted on Figure 1, have been deflated by the CPI index and winsorized at the 1% and 99% levels (computed for each quarter separately). Panels (a) and (b) show quarterly time-series that have been computed by taking equal-weighted averages, which are then logged and HP-filtered. The two panels report correlation coefficients of markups and markdowns with respect to GDP growth; both values (negative 0.49 and negative 0.48) are statistically significant at the 1% level. Panels (c) and (d) show binned scatter plots with 100 equally-sized bins, with (log) real assets on the x-axis and markups/markdowns in level on the y-axis. Dependent and independent variables have been residualized from the quarter fixed effect.

The Return on Loan (RoL) variable, which is displayed on Figure 2, is constructed as the ratio of interest income on loans (RIAD4010) to total loans (RCFD2122). We replace any missing values of total loans with loans net of unearned income and loss allowance. We drop all observations with the level of RoL less than zero. RoL growth is constructed by log-differencing at the bank level. The variable has been winsorized at the 1% and 99% levels (computed for each quarter separately). Expansions and recessions are defined by the NBER criterion. On Panel (b), time series of the mean and skewness of RoL growth are computed by computing the quarterly unweighted average and unweighted statistical skewness, respectively. The series have been first HP-filtered and then run through a moving-average filter with four lags (quarters). Correlation coefficients of the resulting objects with GDP (which has been logged and HP-filtered) are 0.56 and 0.54, respectively,

and both statistically significant at the 1% confidence level.

In Figure 3 the variable on the y-axis of Panel (a) is total assets. Similar results obtain if we use total loans: the GINI coefficients in 1984:1 and 2020:1 were 0.85 and 0.93, respectively. The variable on the y-axis of Panel (b) is total domestic deposits (RCON2200).

Table A1: Variable Details and Sources

Variable	Details	Source
GDP	U.S. real Gross Domestic Product, chained 2012 dollars	FRED (GDPC1)
Consumption	GDP minus Real Gross Private Domestic Investment	FRED (GDPC1-GPDIC1)
Hours	Nonfarm business sector: hours worked for all workers	FRED (HOANBS)
Inflation	Consumer price index for all urban consumers: all items in U.S. city average	FRED (CPIAUCSL)
Assets	Total assets of U.S. commercial banks	Call Reports (RCFD2170)
Loans	Total loans of U.S. commercial banks	Call Reports (RCFD2122)
Equity	Total equity of U.S. commercial banks	Call Reports (RCFD3210)
Deposits	Total domestic deposits of U.S. commercial banks	Call Reports (RCON2200)
Interest income on loans	Total interest income on loans and leases of U.S. commercial banks	Call Reports (RIAD4010)
Loan markups	Estimation procedure is detailed in Appendix A.2	Authors' calculation
Deposit markdowns	Estimation procedure is detailed in Appendix A.2	Authors' calculation
Interest expense	Bank interest expenses on domestic deposits	Call Reports (RIAD4170-RIAD4172)
Expenses	Bank interest and non-interest expenses	Call Reports (RIAD4073+RIAD4093)
Non-interest expense	Bank non-interest expenses	Call Reports (RIAD4093)
Staff cost	Bank expenses on staff	Call Reports (RIAD4135)
Securities	Bank holdings of securities	Call Reports (RCFD1754+RCFD1773)
Non-interest income	Bank non-interest income	Call Reports (RIAD4079)
Fed Funds	Bank holdings of Federal Funds and repos	Call Reports (RCFD3365)
Fed Funds income	Interest income on Federal Funds and repos	Call Reports (RIAD4020)
Fed Funds expense	Interest expense on Federal Funds and repos	Call Reports (RIAD4180)
U.S. Treasuries	Bank holdings of Treasuries and agency debt	Call Reports (RCFDB558)
Income on U.S. Treasuries	Interest income on Treasuries and agency debt holdings	Call Reports (RIADB488)
Deposits charge	Service charges on domestic deposits	Call Reports (RIAD4080)
Net income	Net income of commercial banks	Call Reports (RIAD4340)

Notes: This table summarizes every empirical series used throughout the paper.

Data for Model Calibration We now provide further details on the data that has been used for model calibration. The steady-state level of hours (0.3) represents the share of non-sleeping time that labor market participants in the U.S. spend on working according to the American Time Use Survey (ATUS). It is also a usual value used in the literature (Lee et al., 2020). Average loan markup and deposit markdown targets have been computed by taking an unweighted average of the quarterly asset-weighted averages of the pooled distributions of the estimated $\mu^k_{j,t}$ and $\mu^b_{j,t}$. The corresponding values in the model are average markups $\mu^k(j)$ and markdowns $\mu^b(j)$, weighted by assets $k(j)$.

Average commercial bank leverage has been computed by taking the unweighted average across all banks and quarters of the ratio of total loans to total equity. We discard the vales of leverage below 100. Markup and markdown elasticities of bank size have been estimated by running panel regressions of (log) markups or markdowns on (log) real total bank assets. For these regressions, both markups and markdowns have been first cleaned from the bank-specific averages. Panel regressions include quarter fixed effects. Commercial bank assets and deposits GINI coefficient targets are computed for 2020:1.

Standard deviation of output growth (σ_Y) has been computed from the logged and HP-filtered U.S. real GDP. Consumption data is computed as the difference between real GDP and real private gross investment, following Nuno and Thomas (2017). Labor data is the total hours worked for all workers in the nonfarm business sector (HOANBS), from FRED. All correlation coefficients have been computed on pairs of variables that have been logged and HP-filtered. Correlation coefficient $\rho_{K,Y}$ is computed for the pair of output and bank loans. The latter is our standard variable from the Call Reports. Coefficient $\rho_{K,Y}$ measures the correlation between output and bank equity. $\rho_{LEV,Y}$ is the correlation coefficient between output and book leverage, defined as total loans divided by total equity. Finally, $\rho_{\mu^k,Y}$ $\rho_{\mu^b,Y}$ are correlation coefficients for loan markups and deposit markdowns, respectively.

A.2 Credit Markup and Deposit Markdown Estimation

Loan Markups This section describes how we estimate loan markups and deposit markdowns from U.S. bank-level data. This section follows closely the procedure and description in Bellifemine et al. (2022). We begin with markups, which we define for bank i in quarter t as follows:

$$\mu_{j,t}^{k} = \frac{p_{j,t}}{c_{j,t}}$$

where $p_{j,t}$ is realized interest income on loans and leases divided by total loans and leases, and $c_{j,t}$ is defined as the sum of the ratio of realized interest expenses on domestic deposits and Fed Funds over total deposits and Fed Funds plus marginal net non-interest expenses. Marginal net non-interest expenses are constructed as marginal non-interest expenses minus marginal non-interest income. We estimate marginal non-interest expenses with a trans-log panel fixed-effects regression:

$$\begin{aligned}\log(NIE_{j,t}) = \beta_i + \beta_t + \beta_{l,1}\log(l_{j,t}) + \beta_{w,1}\log(w_{j,t}) + \beta_{q,1}\log(q_{j,t}) \\ + \beta_{l,2}\log(l_{j,t})^2 + \beta_{w,2}\log(w_{j,t})^2 + \beta_{q,2}\log(q_{j,t})^2 + \beta_{l,w}\log(l_{j,t})\log(w_{j,t}) \\ + \beta_{l,q}\log(l_{j,t})\log(q_{j,t}) + \beta_{w,q}\log(w_{j,t})\log(q_{j,t}) + \epsilon_{j,t}\end{aligned} \tag{A1}$$

where $NIE_{j,t}$ is non-interest expenses, β_i and β_t are bank and time fixed effects, respectively, total loans and leases are denoted by $l_{j,t}$, $w_{j,t}$ is staff expenses, computed as the ratio of salaries over assets, and $q_{j,t}$ is total holdings of securities. Further details on variables used are provided in Table A1. From (A1) we obtain *marginal* non-interest expenses as the partial derivative of non-interest expenses with respect to loans:

$$MNIE_{j,t} \equiv \frac{\partial \log(NIE_{j,t})}{\partial \log(l_{j,t}} = \frac{NIE_{j,t}}{l_{j,t}}\left[\beta_{l,1} + 2\beta_{l,2}\log(l_{j,t}) + \beta_{l,w}\log(w_{j,t}) + \beta_{l,q}\log(q_{j,t})\right]$$

Marginal non-interest income estimation follows a similar procedure. This time, however, we do not include inputs into the list of regressors and the dependent variable is now (log) non-interest income:

$$\begin{aligned}\log(NII_{j,t}) = \beta_i + \beta_t + \beta_{l,1}\log(l_{j,t}) + \beta_{q,1}\log(q_{j,t}) + \beta_{l,2}\log(l_{j,t})^2 \\ + \beta_{q,2}\log(q_{j,t})^2 + \beta_{l,q}\log(l_{j,t})\log(q_{j,t}) + \epsilon_{j,t}\end{aligned}$$

As before, marginal non-interest income is defined as the derivative of non-interest income with respect to loans:

$$MNII_{j,t} \equiv \frac{\partial \log(NII_{j,t})}{\partial \log(l_{j,t})} = \frac{NII_{j,t}}{l_{j,t}}\left[\beta_{l,1} + 2\beta_{l,2}\log(l_{j,t}) + \beta_{l,q}\log(q_{j,t})\right]$$

Finally, we define marginal *net* non-interest expenses, *MNNIE* as the difference between marginal non-interest expenses and marginal non-interest income:

$$MNNIE_{j,t} = MNIE_{j,t} - MNII_{j,t}$$

Deposit Markdowns We now proceed with deposit markdown estimation by following a similar recipe as before. We define a markdown for bank j in quarter t as:

$$\mu^b_{j,t} = \frac{c_{j,t}}{p_{j,t}}$$

where $p_{j,t}$ is now a proxy for "safe revenue" that bank j collects in period t and $c_{j,t}$ now represents the marginal cost that bank j must incur in order to raise an extra unit of deposits and maintain the franchise. We measure $p_{j,t}$ as the ratio of realized interest income from Federal Funds, U.S. Treasuries, and agency debt holdings divided by total Fed Funds, U.S. Treasuries, and agency debt holdings. As before, $c_{j,t}$ is now defined as the sum of two objects: the ratio of interest expenses on domestic deposits (net of service charges on domestic deposits) over total domestic deposits, plus marginal net non-interest expenses.

We compute marginal net non-interest expenses as marginal non-interest expenses minus marginal non-interest income. We estimate marginal non-interest expenses with a trans-log panel fixed-effects regression, which is very similar in nature to the one we used for credit markups:

$$\begin{aligned}\log(NIE_{j,t}) = \alpha_i + \alpha_t &+ \beta_{l,1}\log(l_{j,t}) + \beta_{w,1}\log(w_{j,t}) + \beta_{q,1}\log(q_{j,t}) + \beta_{d,1}\log(d_{j,t}) \\ &+ \beta_{l,2}\log(l_{j,t})^2 + \beta_{w,2}\log(w_{j,t})^2 + \beta_{q,2}\log(q_{j,t})^2 + \beta_{d,2}\log(d_{j,t})^2 \\ &+ \beta_{l,w}\log(l_{j,t})\log(w_{j,t}) + \beta_{l,q}\log(l_{j,t})\log(q_{j,t}) + \beta_{w,q}\log(w_{j,t})\log(q_{j,t}) \\ &+ \beta_{l,d}\log(l_{j,t})\log(d_{j,t}) + \beta_{w,d}\log(w_{j,t})\log(d_{j,t}) + \beta_{q,d}\log(q_{j,t})\log(d_{j,t}) + \varepsilon_{j,t}\end{aligned} \tag{A2}$$

where $d_{j,t}$ denotes total domestic deposits,[1] while the definition of all other variables is the same as before.

From (A2) it is straightforward to obtain marginal non-interest expenses as the derivative of non-interest expenses with respect to deposits:

$$MNIE_{j,t} = \frac{\partial \log(NIE_{j,t})}{\partial \log(d_{j,t})} = \frac{NIE_{j,t}}{d_{j,t}}\left[\beta_{d,1} + 2\beta_{d,2}\log(d_{j,t}) + \beta_{l,d}\log(l_{j,t}) + \beta_{w,d}\log(w_{j,t}) + \beta_{q,d}\log(q_{j,t})\right]$$

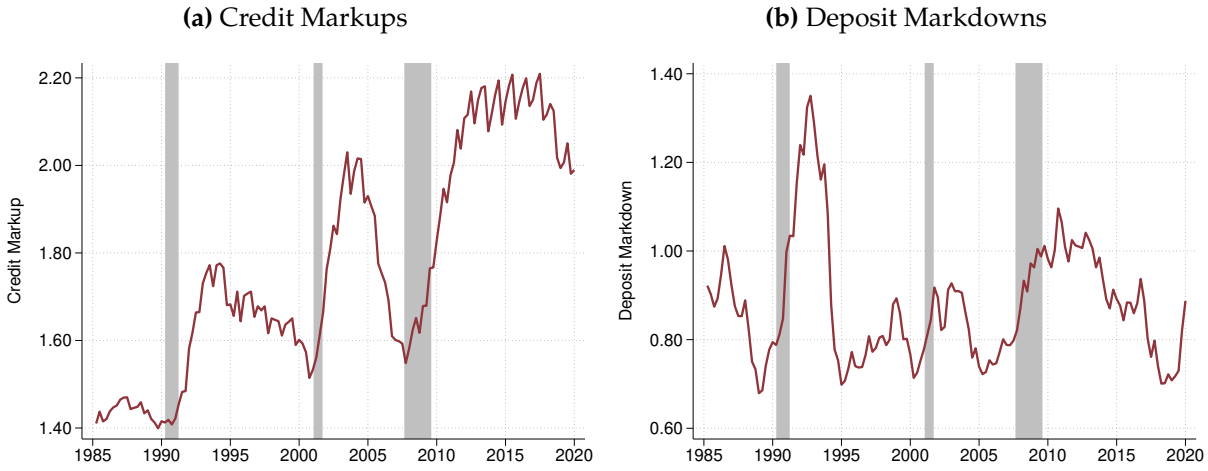
Estimation of marginal non-interest income relies on the exact same procedure, with the usual caveat that we drop inputs from the right-hand side of the regression and the

[1] Like us, Fries and Taci (2005) also use both deposits and loans as proxies for bank-level output.

dependent variable is now (log) non-interest income:

$$\begin{aligned}\log(NII_{j,t}) = \alpha_i + \alpha_t + \beta_{l,1}\log(l_{j,t}) + \beta_{q,1}\log(q_{j,t}) + \beta_{d,1}\log(d_{j,t}) \\ + \beta_{l,2}\log(l_{j,t})^2 + \beta_{q,2}\log(q_{j,t})^2 + \beta_{d,2}\log(d_{j,t})^2 \\ + \beta_{l,q}\log(l_{j,t})\log(q_{j,t}) + \beta_{l,d}\log(l_{j,t})\log(d_{j,t}) + \beta_{q,d}\log(q_{j,t})\log(d_{j,t}) + \varepsilon_{j,t}\end{aligned}$$

Figure A1: Loan Markups and Deposit Markdowns



Notes: This figure plots time series of loan markups and deposit markdowns, computed as quarterly unweighted averages.

Marginal non-interest income is defined as the derivative of non-interest income with respect to deposits:

$$MNII_{j,t} = \frac{\partial \log(NII_{j,t})}{\partial \log(d_{j,t})} = \frac{NII_{j,t}}{d_{j,t}}\left[\beta_{d,1} + 2\beta_{d,2}\log(d_{j,t}) + \beta_{l,d}\log(l_{j,t}) + \beta_{q,d}\log(q_{j,t})\right]$$

Finally, marginal *net* non-interest expenses, $MNNIE$, are computed, analogously to before, as the difference between marginal non-interest expenses and marginal non-interest income:

$$MNNIE_{j,t} = MNIE_{j,t} - MNII_{j,t}$$

Figure A1 plots the resulting estimated time series of μ_t^k and μ_t^b, computed as quarterly unweighted averages of $\mu_{j,t}^k$ and $\mu_{j,t}^b$, respectively. Note that μ_t^k and μ_t^b are reported in levels, unlike in Figure 1 where they are HP-filtered. Three observations are apparent from Figure A1. First, loan markups have been trending up consistently over the past

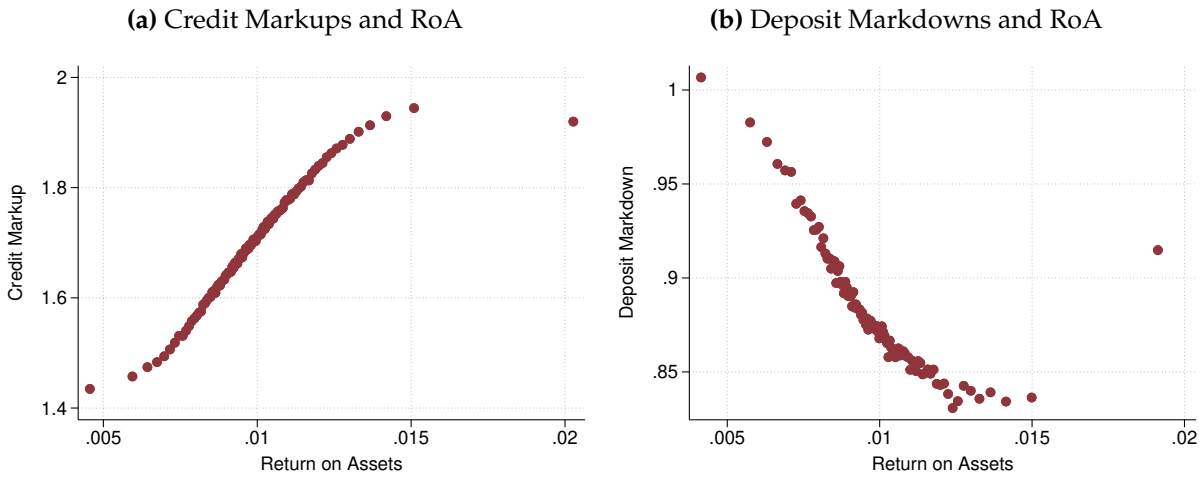
decade, reaching values of above 2 (in gross terms) by 2020. The rise of markups is concentrated around crisis episodes, which is consistent with their counter-cyclicality as we documented before.

Second, deposit markdowns do not exhibit any clear time-series trend. Instead, μ_t^b as a first-order approximation is stationary and centered around 0.85. The markdown is also visibly counter-cyclical, in terms of its unconditional behavior. The markdown is most of the time below unity, implying presence of deposit market power. In early 1990s and early 2010s, μ_t^b climbed to levels of above unity. This is consistent with the spread between the deposit rate and the risk-free rate vanishing to zero during the same episodes (Drechsler et al., 2017).

A.3 Additional Empirical Results

This section presents additional empirical results that supplement our findings in the main text. We begin with a validation check of our measures of loan markups and deposit markdowns. A viable proxy of market power should correlate with measures of profitability. That is, banks that charge high markups or low markdowns should, on average, earn more for the same unit of assets, everything else equal. We test this idea directly by computing bank Return on Assets (RoA) as a ratio of net income over total assets. Figure A2 shows binned scatter plots of RoA on the x-axis and markups (markdowns) on the y-axes of Panel A (Panel B). For both panels, it is clear that market power correlates strongly with higher profits. This result gives more credence to our estimation procedure.

Figure A2: Bank Market Power and Profitability

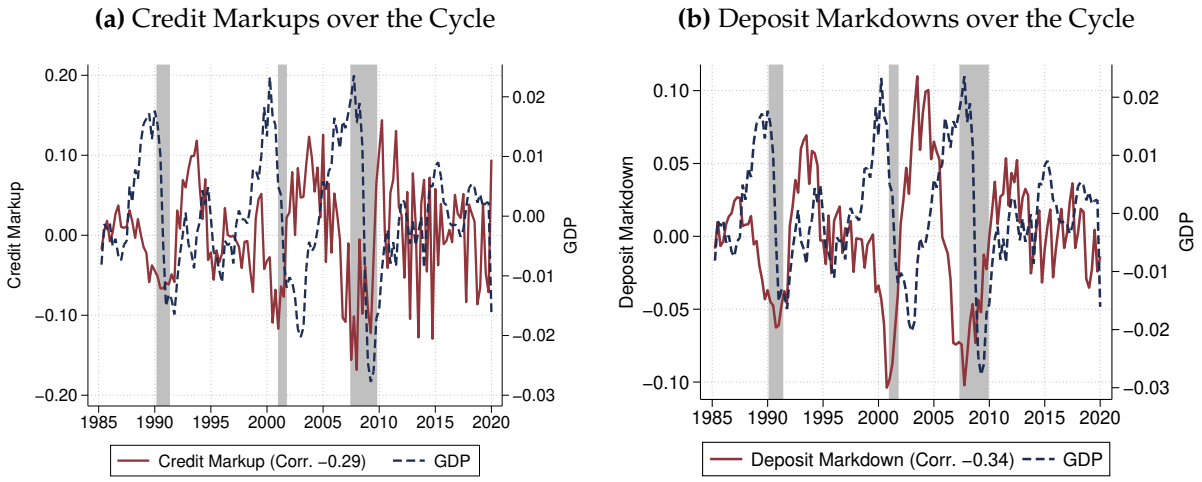


Notes: This Figure plots binned scatter plots of credit markups and deposit markdowns on the y-axes and Return on Assets on the x-axis. Variables have been residualized from the time fixed effect. Each panel features 100 equally-sized bins.

Next, we present two tests of robustness. In main text, our reported quarterly series of markups and markdowns are unweighted averages of bank-level distributions. It is known in the literature that aggregate properties of markups could be affected by how aggregation is performed. We now compute size-weighted average quarterly series μ_t^k and μ_t^b with total assets as a proxy for bank size. As usual, we HP-filter (logged) μ_t^k and μ_t^b and report them alongside U.S. GDP in Figure A3. Cyclical behavior of μ_t^k and μ_t^b is preserved: both are still counter-cyclical with pairwise correlation coefficients with respect to output equal to -0.29 and -.34, respectively, and statistically significant at the 1% level. Size-weighted aggregation therefore mutes the cyclicality of markups and markdowns over the cycle, which is consistent with the idea of market power of large banks (such as,

e.g. the deposit franchise) being more sticky or stable across time.

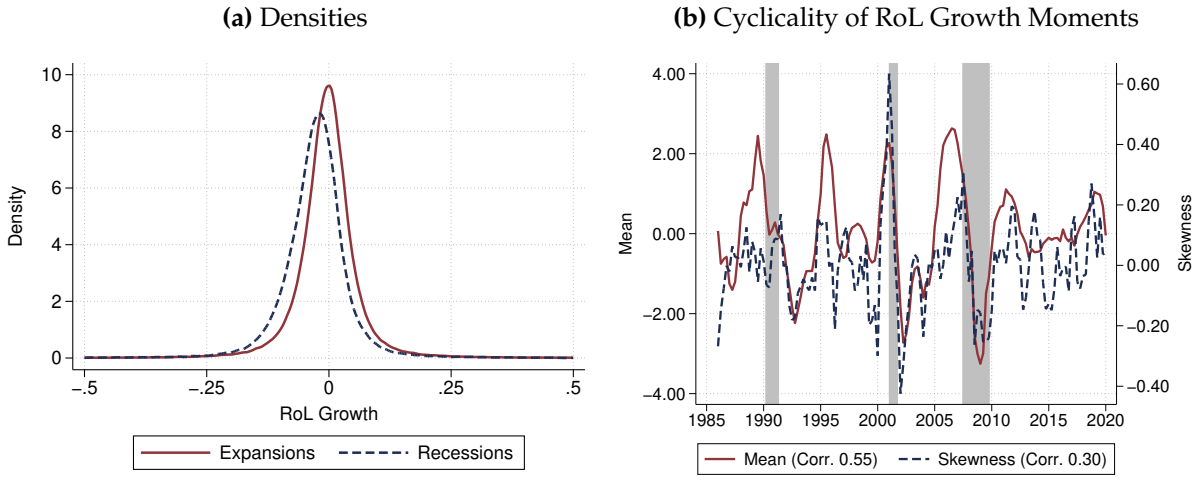
Figure A3: Bank Market Power - Size-Weighted Averaging



Notes: This Figure plots time series of loan markups and deposit markdowns, which have been computed with weighted quarterly averaging. Bank-level total assets are used as weights.

The second check of robustness involves our other key empirical result: counter-cylicality of loan income risk. In main text, we document that the first and third moments of the distribution of bank-level quarterly return on loan (RoL) growth are heavily pro-cyclical. The level of aggregation in that finding was an individual bank. There is a concern that this result would not hold at the level of bank holding companies. We check whether this is the case by re-doing the exercise for the bank holding company level. Figure A4 reports the densities of holding-level RoL growth in recessions and expansions (on Panel A) and time series of the unweighted mean and skewness (on Panel B). The basic takeaway from this test is that the level of aggregation does not influence our results. Bank income risk is counter-cyclical, driven by expansions of the left tail (greater downside risk) in recession. Correlation coefficients of mean and skewness of holding-level RoL growth with U.S. GDP are 0.55 and 0.30, respectively, both statistically significant at the 1%.

Figure A4: Loan Income Risk of Bank Holding Companies



Notes: This figure reports counter-cyclicality of bank income risk at the level of bank holding companies.

B Model Appendix

B.1 Model Details

Household Problem The representative household solves an intertemporal constrained maximization problem subject to a sequence of flow budget constraints and the deposit aggregator. The choice variables in every period are consumption C_t, labor hours L_t, bank-specific deposit savings $\int b_t(j)$, and savings in the mutual fund M_t. The problem takes the following form:

$$\max \mathbb{E}_t \sum_{j=0}^{\infty} \beta^j U\left(C_{t+j}, L_{t+j}, B_{t+j}\right)$$

s.t. the sequence of:

$$C_t + \int_0^1 b_t(j)dj + M_t \leq R_t M_{t-1} + \int_0^1 R_t^b(j) b_{t-1}(j) + L_t W_t + \text{Div}_t + T_t$$

$$B_t = \left[\int_0^1 b_t(j)^{\frac{\theta_b+1}{\theta_b}} dj\right]^{\frac{\theta_b}{\theta_b+1}}$$

The flow utility function is non-separable in hours and separable in deposit holdings: $U(C_t, L_t, B_t) = \frac{1}{1-\phi}\left(C_t - \chi_1 \frac{L_t^{1+\chi_2}}{1+\chi_2}\right)^{1-\phi} + \nu_1 \frac{B_t^{1-\nu_2}}{1-\nu_2}$. The first-order condition with respect labor hours is standard. Note that the non-separability assumption ensures that the marginal utility of consumption does not enter the equation and, thus, the labor supply block can be determined independently from the savings block:

$$L_t = \left(\frac{W_t}{\chi_1}\right)^{-\frac{1}{\chi_2}} \tag{B1}$$

The first-order condition with respect to M_t yields a condition that pins down the risk-free rate:

$$\beta \mathbb{E}_t \left[\frac{U_{C,t+1}(C_{t+1}, L_{t+1}, B_{t+1})}{U_{C,t}(C_t, L_t, B_t)}\right] = \frac{1}{R_{t+1}}$$

where $U_{C,t}(C_t, L_t, B_t) \equiv \left(C_t - \chi_1 \frac{L_t^{1+\chi_2}}{1+\chi_2}\right)^{-\phi}$. The first-order condition with respect to $b_t(j)$ yields:

$$U_{B,t}(C_t, L_t, B_t) \frac{\theta_b}{\theta_b+1} \left(\frac{b_t(j)}{B_t}\right)^{\frac{1}{\theta_b}} \frac{\theta_b+1}{\theta_b} - U_{C,t}(C_t, L_t, B_t) + \beta U_{C,t+1}(C_{t+1}, L_{t+1}, B_{t+1}) R_{t+1}^b = 0$$

where $U_{B,t}(C_t, L_t, B_t) \equiv \nu_1 B_t^{-\nu_2}$. Simple algebra then delivers Equation (B2) in main text:

$$R^b_{t+1}(j) = R_{t+1}\left(1 - \left[\frac{U_{B,t}\left(C_t, L_t, B_t\right)}{U_{C,t}\left(C_t, L_t, B_t\right)}\left(\frac{b_t(j)}{B_t}\right)^{\frac{1}{\theta_b}}\right]\right) \tag{B2}$$

Bank Problem We now derive the Lerner decomposition for the price of claims. Each bank solves:

$$\max V(\mathbf{s};\mathbf{S}) = \mathbb{E}_{\mathbf{s},\mathbf{S}}\{\Lambda'(\mathbf{S'},\mathbf{S})\left[(1-\sigma)n' + \sigma V'(\mathbf{s'};\mathbf{S'})\right]\}$$

subject to:

$$\begin{aligned} n' &= \mathbf{E}_S R^{T'}(\mathbf{s'};\mathbf{S'})qk - R^{b'}b - \zeta_1 k^{\zeta_2} \\ b + n &= k \\ \lambda k &\leq V(\mathbf{s};\mathbf{S}) \end{aligned}$$

as well as the laws of motion of the distribution and stochastic processes. The value function can be simplified in several steps. First, we define an augmented stochastic discount factor as:

$$\tilde{\Lambda}(\mathbf{s'};\mathbf{S'}) \equiv \mathbb{E}_{\mathbf{s},\mathbf{S}}\{\Lambda'(\mathbf{S'},\mathbf{S})\left[(1-\sigma) + \sigma V'(\mathbf{s'};\mathbf{S'})\right]\} \tag{B3}$$

Define $\tilde{\lambda}(\mathbf{s};\mathbf{S})$ as the Lagrange multiplier with respect to the leverage constraint. Plugging in the balance sheet constraint into the law of motion of net worth grants a simplified value function:

$$V(\mathbf{s};\mathbf{S}) = \mathbb{E}_{\mathbf{s},\mathbf{S}}\tilde{\Lambda}(\mathbf{s'};\mathbf{S'})R^{T'}(\mathbf{s'};\mathbf{S'})qk - R^b\left(k-n\right) - \zeta_1 k^{\zeta_2} + \tilde{\lambda}(\mathbf{s};\mathbf{S})\left(V(\mathbf{s};\mathbf{S}) - \lambda k\right) \tag{B4}$$

The first-order condition of Equation (B4) with respect to the choice of claims k yields:

$$\mathbb{E}_{\mathbf{s},\mathbf{S}}\tilde{\Lambda}(\mathbf{s'};\mathbf{S'})R^{T'}\frac{\varepsilon(\mathbf{s};\mathbf{S})-1}{\varepsilon(\mathbf{s};\mathbf{S})}q - R^b - \tilde{\lambda}\left(\mathbf{s};\mathbf{S}\right)\lambda = \zeta_1\zeta_2 k^{\zeta_2-1} \tag{B5}$$

where $\varepsilon(\mathbf{s};\mathbf{S})$ is the credit demand elasticity. Re-writing yields a Lerner decomposition for the price of claims, as in main text:

$$q = \frac{\varepsilon(\mathbf{s};\mathbf{S})}{\varepsilon(\mathbf{s};\mathbf{S})-1}\frac{R^b + \tilde{\lambda}\left(\mathbf{s};\mathbf{S}\right)\lambda + \zeta_1\zeta_2 k^{\zeta_2-1}}{\mathbb{E}_{\mathbf{s},\mathbf{S}}\tilde{\Lambda}(\mathbf{s'};\mathbf{S'})R^{T'}} \tag{B6}$$

Firm Problem The representative capital good producing firm solves a zero-profit, cost minimization static problem every period:

$$\min q_t(j)k_t(j)$$

subject to:

$$\int_0^1 \left[(k_t(j) - \tilde{\gamma}_t)^{\frac{\theta_k - 1}{\theta_k}} \, dj \right]^{\frac{\theta_k}{\theta_k - 1}} = K_t$$

where $\tilde{\gamma} \equiv \gamma_1 K_t^{\gamma_2}$. The inverse demand function is:

$$k_t(j) = \left(\frac{q_t(j)}{Q_t} \right)^{-\theta_k} K_t + \tilde{\gamma}_t \tag{B7}$$

where Q_t is the aggregate price index. We can re-write it as:

$$\frac{q_t(j)}{Q_t} = (k_t(j) - \tilde{\gamma}_t)^{-\frac{1}{\theta_k}} K^{\frac{1}{\theta_k}} \tag{B8}$$

Define the credit demand elasticity as follows:

$$\varepsilon_t(j) \equiv - \frac{\frac{q_t(j)}{Q_t}}{k_t(j) \frac{\partial \left(\frac{q_t(j)}{Q_t} \right)}{\partial k_t(j)}} \tag{B9}$$

The derivative term equals $\frac{\partial \left(\frac{q_t(j)}{Q_t} \right)}{\partial k_t(j)} = -\frac{1}{\theta_k} (k_t(j) - \tilde{\gamma}_t)^{\frac{-1}{\theta_k} - 1} K^{\frac{1}{\theta_k}}$. Substitution and algebra yields:

$$\varepsilon_t(j) = \theta_k \frac{k_t(j) - \tilde{\gamma}_t}{k_t(j)} \tag{B10}$$

Plugging (B10) into (B6) yields the desired equation.

B.2 Additional Results

This section reports additional model-based results which supplement main text. **DISCUSS GRANULAR CRISES**. Table B1 summarizes business cycle fluctuations across all specifications discussed in the paper. The baseline model (column (1)) features both permanent and transitory heterogeneity as well as both credit and deposit market power. The baseline model has been calibrated to match select empirical targets, as per Table (1). In each subsequent column from (2) to (4) we report standard deviations - relative to column (1) - of select macroeconomic and financial variables, obtained from a 5,000 period simulation of each corresponding model specification. In column (6) we report standard deviations relative to the low capital-requirement specification in column (5).

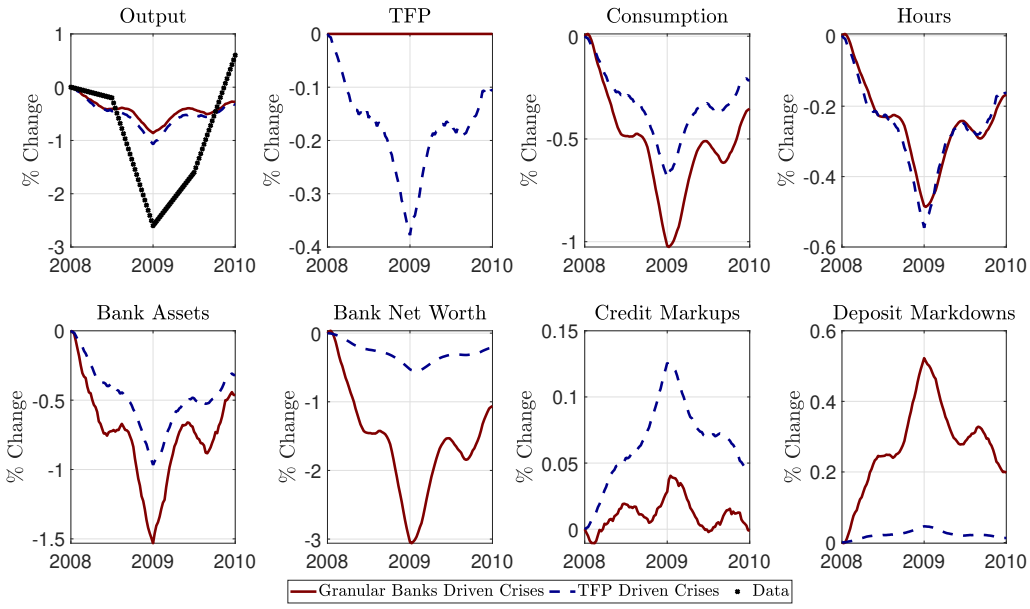


Table B1: Business Cycle Fluctuations: All Specifications

	Relative to Baseline				Relative to Low CR	
	(1)	(2)	(3)	(4)	(5)	(6)
Macro Aggregate	Baseline Model	Only TFP Shocks	Idiosyncratic Shocks to Large Banks	Representative Bank	Low Capital Requirements	High Capital Requirements
Output, Y_t	1	0.65	0.46	0.57	1	0.84
Consumption, C_t	1	0.55	0.62	0.49	1	0.78
Hours, L_t	1	0.73	0.45	0.61	1	0.85
Bank Assets, K_t	1	0.36	0.67	0.27	1	0.76
Bank Net Worth, N_t	1	0.13	0.85	0.10	1	0.60
Bank Deposits, D_t	1	0.40	0.64	0.29	1	0.84
Credit Rates, Q_t	1	0.40	1.14	0.09	1	0.66
Deposit Rates, R_t^b	1	0.98	0.42	0.26	1	0.81
Credit Markups, $\mu_{k,t}$	1	0.51	0.65	0.37	1	0.68
Deposit Markdowns, $\mu_{b,t}$	1	0.09	0.98	0.07	1	0.71

Notes: This table summarizes model-implied business cycle fluctuations across tested specifications. Columns report standard deviations of variable from model simulations.

Three points are worth highlighting. First, shutting down either counter-cyclical idiosyncratic shocks (column (2)) or bank heterogeneity (column (4)) considerably reduces macroeconomic volatility. Second, idiosyncratic shocks to large banks (column (3)) - defined as banks with in the highest quintile of the $\kappa(j)$ distribution of permanent income - by themselves can account for roughly half of business cycle fluctuations. Third and finally, doubling capital requirements from $\lambda = 0.12$ to $\lambda = 0.24$ - conditional on the leverage constraint always binding - reduces business cycle fluctuations by about 16%-22% (column (6)).

B.3 Numerical Methods

This section provides details on the computational algorithm that we use to solve our model. It is a variant of the canonical approach developed by Krusell and Smith (1996, 1998). In order to eventually solve the baseline model with aggregate uncertainty, we proceed with the following these steps:

I. *Solve a simpler model without aggregate uncertainty.*

The full model features an endogenous aggregate state variable N, for which we need to construct an exogenous grid. Before doing so, it is useful to solve for a *stationary equilibrium* and determine the steady-state value of N^{ss}. Specifically, we normalize the value of aggregate productivity A to unity and solve a version of the model that resembles the Aiyagari (1994) framework with endogenous capital accumulation. The special case of our model without aggregate risk is essentially an Aiyagari (1994) model augmented with financial intermediation, which in turn features heterogeneity and imperfect competition.

We solve for the stationary equilibrium on a discrete grid and use interpolation to obtain value and policy function values on off-grid points. For the household problem, we choose an exponential grid for initial deposit holdings b_{-1} with $n_b = 200$ points. We solve the household problem with linear time iteration (Rendahl, 2017). For the bank problem, we set up an exponential grid for net worth $n(j)$ with $n_n = 12$ points. We discretize the distribution of permanent returns $\kappa(j)$ by first drawing a large array of random numbers from a Pareto I density with the shape parameter $\alpha = 1$. The constant κ_m has been normalized such that the distribution's median value is unity. Then, we define 5 permanent return types with the quintiles of the drawn values. We discretize the distribution of left-skewed transitory risk $\xi_t(j)$ with $n_\xi = 5$ nodes. To this end, we use a variant of the Tauchen and Hussey (1991) approach which has been modified to handle the case of non-Gaussian shocks. First, we use the Tauchen and Hussey (1991) method to obtain a matrix of transition probabilities for a stochastic process ξ with volatility σ_ϵ and persistence ρ_ϵ as if it was Normally distributed. Second, we draw a large number of random variables from the Hansen (1994) Skew-t density conditional on the chosen diad $\{\lambda_\epsilon, \eta\}$. The grid for ξ takes on the values of the quintiles of the resulting draw. The median of the grid is normalized to one, and the only difference from the standard case is that our grid is left-skewed. How much more left-skewed it is relative to the Gaussian baseline is controlled by λ_ϵ. We can recover Gaussian nodes under a special case of symmetry, i.e. when $\lambda_\epsilon = 0$. We solve the bank problem by finding a fixed point V^* with value function iteration.

II. *Solving the baseline model with aggregate uncertainty.*

- *Preliminaries.*

  Having solved for the stationary equilibrium, we build an equally-spaced grid with $n_k = 4$ for aggregate bank net worth N, centered around its stationary steady-state value N^{ss}. We assume that aggregate productivity takes on two values: $\{A^h, A^l\}$ with $A^h - A^l = \Delta_a$ chosen as per the calibration table (1) and the discussion surrounding it. The probability matrix that governs transitions across aggregate states is π_a. Now that we have aggregate risk, transitory risk ξ becomes aggregate state-dependent. We assume that high-productivity A^h states are characterized by ξ that is drawn from a Skew-t distribution with $\lambda_\epsilon = 0$ while the low-productivity A^l state features ξ drawn from a Skew-t distribution with $\lambda_\epsilon = -0.5$. That is, banks face Gaussian idiosyncratic shocks in normal states and left-skewed non-Gaussian shocks with a greater downside in bad states.

- *Law of motion of the distribution.*

  A crucial aspect of our numerical solution is how we deal with the endogenous, time-varying distribution of banks μ_t. We follow Krusell and Smith (1998) and assume that banks build limited-information forecasts based on the end-of-period aggregate net worth $N_{t+1} = \int n_{t+1}(j) d\mu_t$ as well as the level of A_t. The limited-information aggregate state vector is thus given by $\mathbf{S}_t = (N_t, A_t)$. We conjecture that the equilibrium mapping Γ is log-linear:

$$A = A^h : \quad \log N_t = \beta_0^h + \beta_1^h \log K_t \tag{B11}$$
$$A = A^l : \quad \log N_t = \beta_0^l + \beta_1^l \log K_t \tag{B12}$$

  The fixed point for Γ is given by the vector $\left(\beta_0^{*h}, \beta_1^{*h}, \beta_0^{*l}, \beta_1^{*l}\right)$.

- *Projection methods.*

  The mapping Γ transitions the distribution of net worth intertemporally. However, banks and the household must also form beliefs over additional aggregate objects intratemporally. Specifically, agents take as given the following aggregate variables: $\left(B_t, Q_t, K_t, W_t, R_t^b, C_t, L_t, \Lambda_t\right)$. We employ linear projection methods and assume and later verify that $\mathbf{S}_t$ is an absorbing aggregate state. That is, once agents know $\mathbf{S}_t$ they can correctly predict every other relevant aggregate. For every relevant aggregate variable X_t in $\left(B_t, Q_t, K_t, W_t, R_t^b, C_t, L_t, \Lambda_t\right)$, we assume that agents form the following projection:

$$A = A^h : \quad \log X_t = \beta_0^{x,h} + \beta_1^{x,h} \log N_t \tag{B13}$$
$$A = A^l : \quad \log X_t = \beta_0^{x,l} + \beta_1^{x,l} \log N_t \tag{B14}$$

  where again note the dependency on the value of A_t. Denote Γ^X the collection of projection rules. Now, let superscript (i) denote an iteration of the algorithm.

We now run the following steps:

(I) *Solve the problem of the banks.*

Start with some initial value for the law of motion of the distribution $\Gamma^{(i)}$ and projections $\Gamma^{X(i)}$. Solve the dynamic banking problem with value function iteration and obtain candidates $\{V, k, b_j, q, R^b\}^{(i)}$.

(II) *Solve the problem of the household.*

Solve the household problem with linear time iteration, conditional on the candidate forecasts $\{\Gamma^B, \Gamma^W, \Gamma^{R^b}\}^{(i)}$. Obtain new candidates for $\{C, L, \Lambda\}^{(i)}$.

(III) *Montecarlo Simulation.*

Simulate the model with a panel of $\mathbf{I} = 1,000$ banks for $\mathbf{T} = 1,000$ periods. In each period, compute the end-of-period aggregate net worth N_t as well as every other aggregate variable using explicit aggregation and without using the forecasting rules.

(IV) *Update laws of motion.*

Using the simulated data, run OLS regressions of (log of) N_t on (log of) N_{t-1} and a constant, having discarded the first 100 periods, and conditional on the aggregate state of the economy $A_t = \{A^h, A^l\}$. Obtain the new candidate for the forecasting rule $\Gamma^{(i+1)}$. Having obtained the sequence of N_t^T, run contemporaneous OLS regressions of (log of) X_t on (log of) N_t and a constant, conditional on the aggregate state of the economy $A_t = \{A^h, A^l\}$, and obtain $\Gamma^{X(i+1)}$.

(V) *Convergence test.*

Compute the Euclidian norm between $\Gamma^{(i)}$ and $\Gamma^{(i+1)}$. If the difference is below a chosen level of tolerance, the algorithm quits. Otherwise, slowly update the forecasting rule as follows: $\Gamma^{(i+1)} = \kappa\Gamma^{(i+1)} + (1-\kappa)\Gamma^{(i)}$ with $\kappa = 0.3$. Similarly, update $\Gamma^{X(i+1)} = \kappa\Gamma^{X(i+1)} + (1-\kappa)\Gamma^{X(i)}$.

Table B2: Equilibrium Forecasting Rules and Accuracy Results

		Γ^*		R^2		Mean Error	SD Error
		A^l	A^h	A^l	A^h	A^l	A^h
N	β_0^*	0.102	0.117	0.995	0.995	0.497%	0.426%
	β_1^*	0.912	0.908				
K	β_0^{K*}	2.259	2.291	0.996	0.995	0.078%	0.060%
	β_1^{K*}	0.585	0.572				
D	β_0^{D*}	2.193	2.232	0.943	0.959	0.120%	0.093%
	β_1^{D*}	0.481	0.467				
Λ	$\beta_0^{\Lambda*}$	-0.005	-0.003	0.985	0.982	0.000%	0.000%
	$\beta_1^{\Lambda*}$	0.000	0.000				
L	β_0^{L*}	-1.368	-1.358	1.000	1.000	0.003%	0.002%
	β_1^{L*}	0.143	0.143				
C	β_0^{C*}	-0.402	-0.392	0.993	0.991	0.006%	0.004%
	β_1^{C*}	0.304	0.304				
W	β_0^{W*}	0.855	0.872	1.000	1.000	0.028%	0.022%
	β_1^{W*}	0.159	0.154				
P	β_0^{K*}	0.114	0.113	0.996	0.973	0.010%	0.007%
	β_1^{K*}	0.023	0.023				
R^b	β_0^{R*}	0.004	0.003	1.000	1.000	0.000%	0.000%
	β_1^{R*}	-0.001	0.000				

Notes: This table reports equilibrium forecasting rules for the law of motion of the distribution as well as every other relevant aggregate variable. It also presents R^2 values from simulation-based linear regressions on equilibrium series. The last two columns report results from the Den Haan (2010) accuracy test and summarize mean and standard deviation of percentage errors between actual and projected simulated series.

Table B2 reports equilibrium forecasting rules obtained as part of the recursive competitive equilibrium solution. The first two columns show the values of β_0^* and β_1^* for every aggregate variable and conditional on low and high aggregate states.

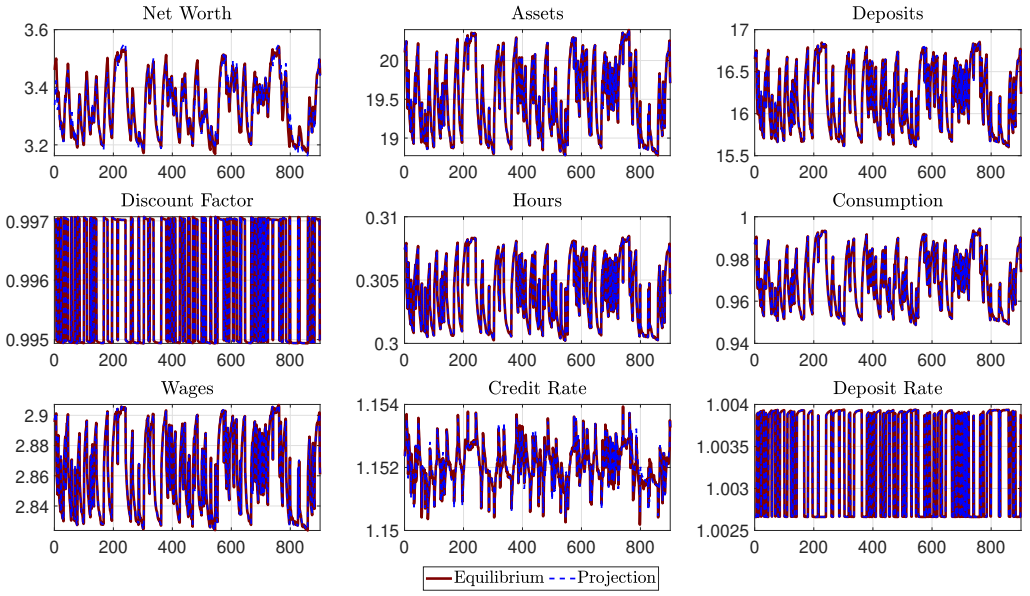
B.4 Solution Accuracy

In order to test whether our algorithm is accurate, we perform two checks. First, it is important to verify that the equilibrium projection rules Γ^* can explain a high percentage of the variation of model-implied aggregates. Table B2 reports the R^2 from the regressions that we run in part (IV) of the numerical algorithm above. The law of motion of the banking distribution is approximated very accurately, as can be seen from the first two

rows, with the R^2 of above 99%. This confirms that approximating the distribution with the first moment does quite well in terms of capturing the dynamics of the full distribution of net worth. We also see that the values of R^2 never drop below 0.94 and are around 0.97 on average across all other aggregates.

A second accuracy test encourages us to go beyond reporting simply the R^2. Den Haan (2010) recommends to compare model-implied time-series of aggregates with forecasts that are built with their corresponding equilibrium forecasting rule Γ^*. Figure B1 plots actual and forecasted values of all the relevant variables. It is clear that projected values track the actual ones very closely. Table B2 reports the mean and standard deviation of the mean percentage difference between actual and projected values for every aggregate. Errors are very low; in particular, mean error for the law of motion of the distribution is less than half of a percentage point with a standard deviation of 0.43%.

Figure B1: Equilibrium and Predicted Aggregates



Notes: This figure plots actual and forecasted series of relevant aggregate variables based on the model simulation.

References

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

Bellifemine, M., R. Jamilov, and T. Monacelli, "HBANK: Monetary Policy with Heterogeneous Banks," *CEPR Working Paper*, 2022, *17129.*

Drechsler, I., A. Savov, and P. Schnabl, "The deposits channel of monetary policy," *Quarterly Journal of Economics*, 2017, *132* (4), 1819–1876.

Fries, Steven and Anita Taci, "Cost efficiency of banks in transition: Evidence from 289 banks in 15 post-communist countries," *Journal of Banking Finance*, 2005, *29* (1), 55–81.

Haan, W. Den, "Assessing the accuracy of the aggregate law of motion in models with heterogeneous agents," *Journal of Economic Dynamics and Control*, 2010, *34*, 79–99.

Hansen, B., "Autoregressive Conditional Density Estimation," *International Economic Review*, 1994, *35*, 705–730.

Krusell, P. and A. Smith, "Income and Wealth Heterogeneity, Portfolio Choice, and Equilibrium Asset Returns," *Macroeconomic Dynamics*, 1996, *1*, 387–422.

_ **and** _ , "Income and Wealth Heterogeneity in the Macroeconomy," *Journal of Political Economy*, 1998, *106*, 867–896.

Lee, S., R. Luetticke, and M. Ravn, "Financial Frictions: Macro vs Micro Volatility," *CEPR DP*, 2020, *15133.*

Nuno, G. and C. Thomas, "Bank Leverage Cycles," *American Economic Journal: Macroeconomics*, 2017, *9(2).*

Rendahl, P., "Linear Time Iteration," *IHS Economics Series Working Paper*, 2017, *330.*

Tauchen, G. and R. Hussey, "Quadrature-Based Methods for Obtaining Approximate Solutions to Nonlinear Asset Pricing Models," *Econometrica*, 1991, *59* (2).