Bewley Banks[†]

Rustam Jamilov **Tommaso Monacelli**

January 2025

Abstract

How do movements in the distributions of bank size and income affect the macroeconomy? To answer this question we develop a dynamic general equilibrium model with heterogeneous financial intermediaries, incomplete markets, and aggregate uncertainty. We find that market incompleteness and uninsured idiosyncratic bank rate of return risk generate minimal concentration in the bank net worth distribution, leading to an "as-if" result, whereby the economy behaves as if it had a representative bank. However, introducing *ex-ante* heterogeneity in the banks' rates of return significantly raises concentration and amplifies real and financial fluctuations relative to the representative-bank case, as this increases a key sufficient statistic, the average marginal propensity to lend. We then extend the model with two empirically-validated features of the banking sector—countercyclical return risk and deposit market power—and show that these amplify and dampen aggregate fluctuations, respectively. Finally, because in the model with ex-ante heterogeneity the distribution of bank size is highly concentrated, shocks to the largest banks can account for almost all of the aggregate variation that is due to idiosyncratic risk, leading to granular banking and economic cycles. The failure of granular banks ("too big to fail") produces sizeable macroeconomic crises.

JEL Codes: E32, E44, G21.

Keywords: Heterogeneous banks; Idiosyncratic bank risk; Incomplete markets; Aggregate uncertainty; Granularity; Bewley models.

[†]We thank the Editor and three anonymous referees for many helpful comments and suggestions that have significantly improved the paper. We thank Dean Corbae, Xavier Gabaix, Mark Gertler, Julien Matheron (discussant), Morten Ravn, Hélène Rey, Joseph Stiglitz, Erwan Quintin, John Vickers 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: All Souls College, University of Oxford. rustam.jamilov@all-souls.ox.ac.uk.

Monacelli: Bocconi University, CEPR, and IGIER. tommaso.monacelli@unibocconi.it.

1 Introduction

The 2007-2008 global financial crisis prompted a significant reassessment of the role financial intermediaries play in the macroeconomy (Gertler and Kiyotaki, 2010; Brunnermeier and Sannikov, 2014). Fifteen years later, the 2023 U.S. regional banking crisis and the collapse of Credit Suisse have once again brought banks into the spotlight for both academics and regulators alike. As such, understanding how banks contribute to business-cycle fluctuations remains a matter of critical economic importance.

Most of the existing studies have so far focused on a *representative* intermediary in environments where the cross-section of banks plays no role.[1] In this paper, we develop a tractable, quantitative macroeconomic framework where a dynamic distribution of banks has first-order effects on aggregate fluctuations. Our starting framework features aggregate uncertainty and *ex-post* bank heterogeneity, which is due to incomplete markets and uninsured idiosyncratic bank rate of return risk. Each financial intermediary invests into risky claims on non-financial firms, sources deposits from households, and is subject to an agency friction. Under perfect insurance, and in the absence of idiosyncratic shocks, our framework nests the canonical Real Business Cycle model and the Gertler and Kiyotaki (2010) and Gertler and Karadi (2011) class of influential representative-bank macro-banking models as special cases.

Our modeling approach eliminates *scale invariance*: all dynamic choices in the financial sector depend on bank-specific net worth. The resulting equilibrium yields a non-trivial, dynamic distribution of bank size. The presence of aggregate risk makes this distribution, in principle, an infinitely-dimensional dynamic object and a relevant state variable. We resort to numerical methods and to the Krusell and Smith (1996, 1998) algorithms to solve the model fully non-linearly. We show that the equilibrium, reminiscent of the result in the stochastic growth model of Krusell and Smith (1998), features "approximate aggregation", whereby all aggregate variables can be accurately described as a function of two states: exogenous aggregate productivity and the first moment of the bank net worth distribution.

The combination of ex-post heterogeneity and aggregate uncertainty generates a bank net worth fluctuation problem analogous to the canonical Bewley-Imrohoroğlu-Huggett-Aiyagari framework (Bewley, 1977; Imrohoroğlu, 1989; Huggett, 1993; Aiyagari, 1994). With these features, we show that the model produces an "as-if" result, whereby the economy behaves as if it were populated by a representative bank (Werning, 2015). In order to understand this finding, we introduce the notion of *marginal propensity to lend*

[1] We discuss the handful of relevant exceptions in the literature portion of this section.

(MPL), defined as the bank-specific response of lending to a marginal change in bank-specific net worth. The key insight is that the general-equilibrium average MPL in the model with ex-post heterogeneity is essentially identical to the MPL of the representative bank. While the distribution of banks features some dispersion and net worth moves around, it is still strongly at odds with the data: there are too few small banks with low levels of net worth and the degree of concentration among the high-net worth banks is too low. Hence, all net worth is accumulated by banks with approximately the same slope of the lending policy function, leading to approximately symmetric lending choices. As a result, exact aggregation is achieved, with the distribution having limited aggregate effects.

Next, we extend the model by introducing a further core element: *ex-ante* heterogeneity in the banks' rate of return. In addition to stochastic shocks, banks now permanently differ in terms of an ex-ante characteristic. This feature could be capturing, for example, permanent differences in the quality of "screening devices" (Stiglitz and Weiss, 1981) across banks. The model with both ex-ante and ex-post heterogeneity generates realistic, right-skewed ergodic distributions of bank size (assets, deposits, and net worth). Both asset and deposit markets are considerably concentrated, almost exclusively due to the presence of permanent bank returns inequality. This version of the model also generates financial and business-cycle statistics that approximate cyclical properties of the different moments of the U.S. economy rather well.

We find that the baseline model with *both* ex-ante and ex-post bank heterogeneity, unlike the starting version with only idiosyncratic risk, significantly *amplifies* real and financial fluctuations relative to the representative-bank case. In this baseline economy, the MPL is heterogeneous, increasing in permanent returns, and declining in bank size. Smaller banks have a greater sensitivity of lending with respect to changes in net worth. Crucially, the average MPL of the economy is greater than the MPL in a representative-bank benchmark. Ex-ante heterogeneity introduces a mass of banks that are very large and with a low MPL. However, their share is not 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 case, and the economy—total bank lending and aggregate output in particular—is more responsive to exogenous shocks.

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 net worth 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.

Granularity plays a crucial role in shaping the business cycle in our model. More than sixty per cent of the variation in aggregate output due to idiosyncratic risk can be accounted for by shocks hitting only the *top decile* of the banking distribution. This pattern is reminiscent of the Pareto principle and the "eighty-twenty" rule (Gabaix, 2009). In our model, 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).

Our model is well-suited for the study of two salient dimensions of financial crises: the ones deriving from large aggregate contractions and the ones deriving from bank failures. For one, we employ the event-study approach that is popular in the open-economy macroeconomics literature (Calvo et al., 2006; Mendoza, 2010), simulate the model for a large number of periods, and identify crisis events as incidents of aggregate output falling below a certain threshold. Our framework generates realistic banking and economic crises. When the baseline economy is parameterized to fit the collapse of U.S. GDP during the Great Recession, the representative-bank economy can account for less than one-third of the actual observed contraction of output in the data. Also, the contraction in bank assets and net worth during a financial crisis episode is an order of magnitude larger in the Bewley Banks economy relative to the representative-bank case. Second, we focus on the macroeconomic consequences of *bank failures*. We study the economic impact of defaults of different bank types (large vs small). The failure of granular banks (i.e., banks that belong to the highest size percentile) leads to large crises: fifty per cent of the economy-wide bank equity is wiped out and aggregate output contracts by seven and a half per cent. Hence, our model features simultaneously granularity and a *too-big-to-fail* property.

Finally, we extend the model with two empirically-validated features of the banking sector: *counter-cyclical* rate of return risk and *deposit market power*. In the data, transitory return risk faced by commercial banks 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. We find that counter-cyclical idiosyncratic bank return risk *per se* amplifies aggregate fluctuations. 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 experiences extremely bad portfolio outcomes. Bank lending contracts, aggregate production stalls, and consumption falls.

Departing from the perfect competition assumption, we finally introduce market power on the deposit side of banks. Households save either in mutual funds or bank deposits, but derive utility from the special liquidity services provided by deposits. Internalizing this effect, banks charge a *markdown* below the risk-free rate. Because marginal liquidity preferences are aggregate state-dependent, markdowns vary over the business cycle. Moreover, since banks are heterogeneous, markdown choices explicitly depend on bank size, the distribution of which is itself aggregate state-dependent. The model generates counter-cyclical deposit markdowns, in line with the data. This result hinges on the endogenous stickiness of the interest rate on deposits. In a bad aggregate state, bank lending and output both contract, but gradually, due the slow response of net worth. Since the contraction is persistent, and follows a hump-shaped dynamic, consumption growth falls, leading to a fall in the risk-free real interest rate. Banks, however, exercise their market power in order to dampen a deposit flight and lower the deposit rate *by less* than the risk-free rate, leading to a counter-cyclical markdown. Overall, deposit market power *dampens* the effects of aggregate shocks relative to the baseline model with perfect competition. Banks, by raising markdowns in bad aggregate states, try to protect the demand for deposits, thereby allowing depositors (households) at the margin to smooth their response to shocks, preventing deposit withdrawals and the resulting contraction in lending.

Related Literature The paper relates to several literature strands that span macroeconomics, banking, and financial economics. To begin with, our work belongs to the new, burgeoning literature on heterogeneous financial intermediaries. 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); Dempsey (2024); Goldstein et al. (2023); Abadi et al. (2023). 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 (2021) build a quantitative model of banking industry dynamics with uninsured idiosyncratic return risk and imperfect credit-market competi-

tion. Bianchi and Bigio (2022) study the credit channel of macroeconomic transmission in a macro-banking framework with stochastic deposit withdrawal shocks. Begenau and Landvoigt (2021) build a quantitative model with two banking sectors that approximate the empirically-documented divide between standard commercial and "shadow" banks. Mendicino et al. (2024) study optimal bank capital requirements in a quantitative model with "twin defaults": simultaneous failures of financial and non-financial firms. Our contribution relative to this stream of papers is to (i) study the interaction of both ex-ante and ex-post bank heterogeneity with aggregate uncertainty, (ii) emphasize the marginal propensity to lend as a sufficient statistic for equilibrium model dynamics, (iii) and extend the baseline framework with counter-cyclical bank income risk and deposit market power—two salient features of the data.

Furthermore, we are contributing to the so-called "macro-finance" section of the literature that embeds the financial intermediary sector into macroeconomic frameworks. (Cúrdia and Woodford, 2001; Brunnermeier and Pedersen, 2009; Adrian and Shin, 2010; Jermann and Quadrini, 2013; He and Krishnamurthy, 2013; Adrian and Shin, 2014; Clerc et al., 2015; Elenev et al., 2021; Bocola and Lorenzoni, 2023; Amador and Bianchi, 2024). 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 Gertler et al. (2012), Bocola (2016), Gertler et al. (2016), Nuno and Thomas (2017), Gertler et al. (2020), and Faccini et al. (2024). Our study particularly emphasizes the departure from the representative intermediary assumption and its first-order impact on business-cycle fluctuations.

Our paper builds on a vast banking literature (Diamond, 1984; Bernanke and Blinder, 1988; Holmstrom and Tirole, 1997; Diamond and Dybvig, 1983; Carlstrom and Fuerst, 1997; Bernanke and Gertler, 1995; Allen and Gale, 1998; Hellman et al., 2000; Allen and Gale, 2004). The counter-cyclical return risk extension of our baseline model is inspired by the studies that find similar patterns for households and non-financial firms and show that those matter for business cycles (Storesletten et al., 2004; Guvenen et al., 2014; Krueger et al., 2016). Finally, the deposit market power extension of our framework is modelled in the spirit of the deposits channel of monetary policy (Drechsler et al., 2017, 2021; Wang et al., 2022) and the money-in-utility framework (Sidrauski, 1967; Galí, 2008). Some of the contributions to the vibrant literature on banking competition include Boyd and Nicolo (2005); Egan et al. (2017); Heider et al. (2019); Kurlat (2019); Polo (2021); Whited et al. (2021); Di Tella and Kurlat (2021); Wang (2024).

2 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 capital good producer, a representative final good producer, and a continuum of heterogeneous financial intermediaries (banks, for short) that are indexed by $j \in [0, 1]$.

2.1 Households

The representative household's preferences are separable inter-temporally and discounted at the rate of $\beta \in (0, 1)$. The household derives utility from consumption, C_t, and disutility from labor (measured in hours), H_t. Preferences are given by:

$$\max \mathbb{E}_t \sum_{k=0}^{\infty} \beta^k U\Big(C_{t+k}, H_{t+k}\Big) \tag{1}$$

Denote with ψ the elasticity of intertemporal substitution, $\frac{1}{\chi_2}$ the elasticity of labor supply, and χ_1 the parameter that gauges labor disutility. The period utility function is of CRRA form and features intra-temporal non-separability between consumption and hours in the spirit of Greenwood et al. (1988):

$$U(C_t, H_t) = \begin{cases} \frac{1}{1-\psi}\left(C_t - \chi_1 \frac{H_t^{1+\chi_2}}{1+\chi_2}\right)^{1-\psi} & , \psi \neq 1 \\ \ln\left(C_t - \chi_1 \frac{H_t^{1+\chi_2}}{1+\chi_2}\right) & , \psi = 1 \end{cases} \tag{2}$$

Households can save in the form of one-period deposits, $b_t(j)$, in bank j. Deposits pay a non-contingent gross rate of return R_t. Let W_t be the wage rate, T_t lump-sum taxes, and Π_t bank dividend transfers, all of which are taken as given. Households maximize utility subject to a sequence of budget constraints:

$$C_t + \int_0^1 b_t(j) dj \leq \int_0^1 R_t b_{t-1}(j) + H_t W_t + \Pi_t + T_t \tag{3}$$

The first-order condition for bank deposits is given by:

$$R_{t+1} = \left[\mathbb{E}_t\left(\frac{\beta u'\Big(C_{t+1}\Big)}{u'\Big(C_t\Big)}\right)\right]^{-1} \tag{4}$$

where $\Lambda_{t+1} \equiv \beta \frac{u'(C_{t+1})}{u'(C_t)}$ denotes the stochastic discount factor. Finally, the first-order condition for labor supply is given by:

$$H_t = \left(\frac{W_t}{\chi_1} \right)^{\frac{1}{\chi_2}} \tag{5}$$

2.2 Non-Financial Firms

There is a continuum of measure one of perfectly competitive firms that produce the final good, Y_t, using an identical constant returns to scale Cobb-Douglas production function with aggregate capital, K_t, and labor, H_t, as inputs:

$$Y_t = A_t K_t^{\alpha} H_t^{1-\alpha}, \quad 0 < \alpha < 1 \tag{6}$$

where A_t is aggregate productivity. A_t is stochastic and takes on two possible values: A_H and A_L which represent, respectively, good and bad states. The shock follows a first-order Markov structure with $\boldsymbol{\pi_a}$ the matrix of transition probabilities.

There is a continuum of identical capital producing firms indexed by i. These firms are cash-strapped and depend on banks for external financing. Individual bank-level claims on firms, $l_t(j)$, are perfect substitutes and get aggregated into the stock of total claims as follows:

$$L_t = \int_0^1 l_t(j) dj \tag{7}$$

Each capital producing firm intakes claims $L_t(i)$ and produces $\Phi(L_t(i))$ new units of the capital good, where $\Phi(.)$ is a transformation function such that $\Phi' > 0$ and $\Phi'' < 0$. The decision problem of each producer is thus:

$$\max_{L_t(i)} Q_t \Phi\Big(L_t(i)\Big) - L_t(i)$$

Given the symmetry for capital producers ($L_t(i) = L_t$), the aggregate price of capital is pinned down as follows: $Q_t = \left[\Phi'\Big(L_t\Big) \right]^{-1}$. Finally, the return on capital—which banks will be taking as given—is given by:

$$R_{t+1}^k = \frac{A_{t+1} \alpha K_{t+1}^{\alpha-1} H_{t+1}^{1-\alpha}}{Q_t} \tag{8}$$

After being used in the production of the final good, capital fully depreciates.

2.3 Banks

There is a continuum of banks indexed by j. The role of banks in the economy is to source deposits, $b_t(j)$, from households and invest into the capital producing firms via claims, $l_t(j)$. Deposits pay a state non-contingent interest rate, R_t. Banks are risk neutral, accumulate net worth, $n_t(j)$, maximize the present discounted value of their franchise value, $V_t(j)$, and exit the economy with an exogenous probability $(1-\sigma) > 0$, upon which their franchise gets transferred to the household in the form of dividends.[2]

Ex-Ante and Ex-Post Heterogeneity Banks are ex-ante heterogeneous due to permanent differences in the efficiency of intermediation $\kappa \in \Theta$. A higher $\kappa(j)$ allows bank j 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 skills and ability (Diamond, 1984) or screening devices (Stiglitz and Weiss, 1981).

In addition, markets are incomplete and bank returns feature uninsured idiosyncratic rate of return risk, $\xi_t(j)$, in the spirit of Benhabib and Bisin (2018) and Benhabib et al. (2019). All banks earn a common aggregate net return on capital, r_t^k, which is perturbed by the aforementioned permanent and transitory components of rate of return heterogeneity. The bank-level gross total portfolio return, $R_t^T(j)$, is therefore:[3]

$$R_t^T(j) = 1 + \kappa(j)\xi_t(j)r_t^k \tag{9}$$

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

$$\kappa(j) \sim P(\alpha_\kappa, k_m), \quad Prob(\kappa > k_m) = \left(\frac{k_m}{\kappa}\right)^{\alpha_\kappa} \tag{10}$$

The transitory rate of return risk, $\xi \in \Xi$, follows an AR(1) process with shocks drawn from a Gaussian distribution:

$$\xi_t(j) = (1-\rho_\xi)\mu_\xi + \rho_\xi \xi_{t-1}(j) + \epsilon_t(j), \quad \epsilon \sim \mathcal{N}(0, \sigma_\xi) \tag{11}$$

[2]Observe that there is no endogenous bank exit and dividends are paid out only once. In contrast, Corbae and D'Erasmo (2021) allow for endogenous exit and pre-exit dividend payments.

[3]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. See also the review article by Kaplan and Violante (2022) with a summary of ex-post and ex-ante heterogeneity approaches in modern macroeconomics.

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: $\zeta_1 l_t(j)^{\zeta_2}$ with $\zeta_1 > 0$ and $\zeta_2 > 1$.[4] Importantly, the 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 be written as:

$$n_{t+1}(j) = R^T_{t+1}(j) l_t(j) - R_{t+1} b_t(j) - \zeta_1 l_t(j)^{\zeta_2} \tag{12}$$

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

$$b_t(j) + n_t(j) = l_t(j) \tag{13}$$

Agency Friction 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 $\lambda > 0$ of the franchise value, $V_t(j)$, for personal use. If a diversion is successful, the franchise is bankrupt and only the remaining fraction, $(1 - \lambda)$, is recovered by depositors. In order to limit the banker's incentive to divert assets, bank assets should not exceed an endogenous threshold given by the franchise value of the bank. This yields the following incentive constraint that limits the leverage multiple:

$$\lambda l_t(j) \leq V_t(j) \tag{14}$$

Dynamic Bank Problem We now adopt a recursive notation and drop the time and (j) indexations temporarily. The aggregate state of the economy is characterized by (Γ, A). Γ 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 cross-sectional distribution of bank net worth, permanent return types, and transitory return draws. The part of the law of motion of the aggregate state that concerns A is exogenous and can be described by the transition matrix $\boldsymbol{\pi}_a$. The law of motion of the distribution is denoted by F, such that $\Gamma' = F(\Gamma, \mathrm{A}, \mathrm{A}')$.

The relevant idiosyncratic state vector 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 by β and augmented by the exogenous dividend rule $(1 - \sigma)$. Banks take their initial net worth as given and choose how much to invest into firms, l, and how much to borrow from households, b. The competitive structure in both asset and

[4] Alternatively, the adjustment of the quantity of loans is costly and the cost must be incurred at the bank franchise level.

deposit markets, exogenous processes, and aggregate prices are taken as given. Taking into account that banks cannot operate with negative equity, the dynamic banking problem takes the following form:

$$V(n,\kappa,\xi;\Gamma,A) = \max_{\{l,b,n'\}\geq 0}\left\{\beta\mathbb{E}\left[(1-\sigma)n' + \sigma V(n',\kappa,\xi';\Gamma',A'|\xi,A)\right]\right\} \tag{15}$$

subject to:

$$\begin{aligned} n' &= \Big[1+\kappa\mathbb{E}\Big(\xi' r^{k'}(\Gamma',A'|\Gamma,A)\Big)\Big]l - R'b - \zeta_1 l^{\zeta_2} \\ b+n &= l \\ \lambda l &\leq V(n,\kappa,\xi;\Gamma,A) \\ \Gamma' &= F(\Gamma,A,A') \end{aligned}$$

Partial Equilibrium We can obtain a partial-equilibrium solution to the banking problem, in which the law of motion Γ and all aggregate quantities and prices are taken as given. First, we substitute out the balance sheet constraint and the law of motion of net worth in order to obtain a simplified version of the problem:

$$V(n,\kappa,\xi;\Gamma,A) = \max_{l\geq 0}\left\{\mathbb{E}\Omega\Big[\Big(1+\kappa\xi' r^{k'}(\Gamma',A'|\Gamma,A) - R'\Big)l - \zeta_1 l^{\zeta_2} + R' n\Big]\right\}$$

subject to:

$$\begin{aligned} \lambda l &\leq V(n,\kappa,\xi;\Gamma,A) \\ \Gamma' &= F(\Gamma,A,A') \end{aligned}$$

where we have defined $\Omega \equiv \beta\Big(1-\sigma+\sigma\frac{V(n',\kappa,\xi';\Gamma',A'|\xi,A)}{n'}\Big)$ as the bankers' augmented discount factor. Risk-neutral banks lever up until the leverage constraint binds. Then, the policy function for banks' choice of assets, $\mathcal{L}$, as an implicit function of the choice variable, l, is:

$$\mathcal{L}(n,\kappa,\xi;\Gamma,A) = \frac{\mathbb{E}\Big\{\Omega\Big(R'n - \zeta_1 l^{\zeta_2}\Big)\Big\}}{\lambda - \mathbb{E}\Big\{\Omega\Big(1+\kappa\xi' r^{k'}(\Gamma',A'|\Gamma,A) - R'\Big)\Big\}} \tag{16}$$

The complete solution to the banking problem is therefore a pair $(V,\mathcal{L})$, i.e., the value function and the corresponding policy function for claims.

Before proceeding, it is useful to briefly compare the formula for $\mathcal{L}$ above with the analogous expression $\mathcal{L}_{GK}$ in the representative-bank benchmark (Gertler and Kiyotaki

(2010), equation 27):

$$\mathcal{L}_{GK} = \frac{\mathbb{E}\left\{\Omega R'\right\}}{\lambda - \mathbb{E}\left\{\Omega\left(1 + r^{k'}(\Gamma', A'|\Gamma, A) - R'\right)\right\}} n$$

There are two substantive differences. First, note the presence of non-interest expenses in the numerator of equation (16). This departure implies that the scale of the bank matters and net worth is a relevant state variable, which is not the case in the standard model. Second, observe the presence of permanent and stochastic heterogeneity, κ and ξ, in equation 16. This suggests that—irrespective of whether the model is scale invariant or not—banks' lending choices will vary by the level of both κ and ξ. Taken together, the two departures confirm that the relevant idiosyncratic state vector of the banking problem is (n, κ, ξ), which is substantively richer than in the standard framework.

Marginal Propensity to Lend It is useful to introduce the concept of *marginal propensity to lend* (MPL), defined as the change in bank-level lending out of a marginal change in bank's net worth. The MPL can be computed directly from equation (16):

$$MPL = \frac{\mathbb{E}\left\{\Omega R'\right\}}{\lambda - \mathbb{E}\left\{\Omega\left(1 + \kappa\xi' r^{k'}(\Gamma', A'|\Gamma, A) - R'\right)\right\} + \zeta_1\zeta_2 l^{\zeta_2 - 1}} \tag{17}$$

It is clear from this formula that the MPL varies by both permanent and transitory returns (κ and ξ, respectively) as well by size (l). Note that the MPL is increasing in κ and ξ and decreasing in size. It is again useful to compare (17) to its representative-bank counterpart in Gertler and Kiyotaki (2010):

$$MPL_{GK} = \frac{\mathbb{E}\left\{\Omega R'\right\}}{\lambda - \mathbb{E}\left\{\Omega\left(1 + r^{k'}(\Gamma', A'|\Gamma, A) - R'\right)\right\}}$$

Notice that the MPL_{GK} in this case does not vary by bank—either by size or by the return profile—and is determined solely by aggregate variables. Conversely, the *sensitivity* of lending to net worth changes will vary by bank in our framework. It is important to emphasize that this still pertains to a partial-equilibrium logic, i.e., when the banks' distribution and aggregate quantities and prices are taken as given. We now describe the general equilibrium properties of the model.

2.4 Dynamics of the Banking Distribution

In order to solve the banking problem above, individual banks must forecast the distribution, Γ. This is necessary in order to determine the next period's return on aggregate capital, R^k_{t+1}, which depends on the future capital stock, K_{t+1}, which is in turn pinned down by claims, L_{t+1}. However, Γ is a high-dimensional object, endogenous, and time-varying. We build on the canonical Krusell and Smith (1998) method and assume that banks form linear, limited-information forecasts based on a small set of I moments of the distribution $\boldsymbol{m} \equiv (\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_I)$. The law of motion can then be reformulated in terms of the dynamics of these moments: $\boldsymbol{m}' = F(\boldsymbol{m}, A, A')$.

Specifically, we choose $I = 1$ and assume that F is log-linear. We later verify that tracking the first moment is sufficient. The forecast for the first moment (mean) of the distribution of net worth, $\bar{N}$, is thus as follows:

$$A = A_L: \quad \log \bar{N}' = \beta^N_{L,0} + \beta^N_{L,1} \bar{N} \tag{18}$$

$$A = A_H: \quad \log \bar{N}' = \beta^N_{H,0} + \beta^N_{H,1} \bar{N} \tag{19}$$

The forecast for the first moment of the distribution of bank claims, $\bar{L}$, conditional on net worth, is:

$$A = A_L: \quad \log \bar{L}' = \beta^L_{L,0} + \beta^L_{L,1} \bar{N} \tag{20}$$

$$A = A_H: \quad \log \bar{L}' = \beta^L_{H,0} + \beta^L_{H,1} \bar{N} \tag{21}$$

Conditional on the forecasted $\bar{L}_{t+1}$, quantities K_{t+1} and H_{t+1}, and prices W_t and Q_t are pinned down by the joint solution to the non-financial firm problem and the labor supply condition. As a result, the forecast for R^k_{t+1} can be successfully constructed and the banking problem can be solved. See Appendix C.1 for details on our computational algorithm. See also Appendix C.2 for numerical accuracy checks.

2.5 Market Clearing and General Equilibrium

In order to close the model, we impose the market clearing conditions. The goods market equilibrium is as follows:[5]

$$Y_t = C_t \tag{22}$$

[5] As is common in the literature, we assume that aggregate non-interest expenses are rebated lump-sum to the household for their financial services.

Equilibrium in the credit market requires:

$$\int_{\xi}\int_{\kappa}\int_{n} n^*(n,\kappa,\xi)\Gamma_{t-1}dnd\kappa d\xi + \int_{\xi}\int_{\kappa}\int_{n} b^*(n,\kappa,\xi)\Gamma_t dnd\kappa d\xi = \int_{\xi}\int_{\kappa}\int_{n} \mathcal{L}(n,\kappa,\xi)\Gamma_t dnd\kappa d\xi \tag{23}$$

where n^*, b^*, and $\mathcal{L}(n,\kappa,\xi)$ are policy functions for net worth, deposit supply, and claims, respectively. Equilibrium in the asset market reads as follows:

$$K_{t+1} = \Phi\left(\int_{\xi}\int_{\kappa}\int_{n} \mathcal{L}(n,\kappa,\xi)\Gamma_t dnd\kappa d\xi\right) \tag{24}$$

Finally, the labor market clears by Walras' law.

Equilibrium A recursive competitive equilibrium consists of the law of motion of the banking distribution Γ, the bank value and policy functions $(V,\mathcal{L})$, and policy functions for the household (C^*,H^*,B^*) such that, given a vector of aggregate pricing functions $\{R,R^k,Q,W\}$, (i) the value and policy functions of all agents solve the corresponding decision problems; (ii) Γ is consistent with agents' optimization; (iii) prices are determined by first-order conditions as above; (iv) all markets clear.

2.6 Discussion

Before proceeding, we briefly discuss some of our modeling choices and implications. First, a critical building block of our model is market incompleteness and uninsured idiosyncratic bank rate of return risk. In the classic banking model of Diamond (1984), idiosyncratic bank income risk is eliminated if banks can fully diversify their lending and investment positions. In practice, however, bank portfolios are often concentrated and unhedged, which comes with financial and real-economy consequences (Galaasen et al., 2021). The object $\xi_t(j)$ thus captures, in a parsimonious and reduced-form way, the banks' exposure to idiosyncratic firm performance shocks and the inability to hedge that risk.

Second, we have assumed that the sole general type of liability that banks can issue to raise funds is short-term non-contingent debt. In reality, state-contingent debt (outside equity) provides an additional hedging instrument against idiosyncratic and aggregate fluctuations in net worth. Gertler et al. (2012) build a macroeconomic model with a representative bank that has the option to issue outside equity and, as a result, navigate its risk exposure. While it is in principle possible to introduce costly outside equity issuance into our baseline model, our results are unlikely to change materially if the cost is sufficiently high and/or convex in the short run. Banks with low levels of net worth—

which, as we will see, are critical for our analysis—would be particularly unlikely to bear such cost, yielding similar distributional implications.

Third, an additional source of external funds for banks can be the wholesale funding (interbank) market. Lending and borrowing on the interbank market can help banks navigate the short-run liquidity and net worth fluctuation problems. Bianchi and Bigio (2022) build a macro-finance framework with a representative intermediary, idiosyncratic deposit withdrawal risk, and an over-the-counter interbank market. They show that frictional interbank trading can have first-order effects on the macroeconomy and the conduct of monetary and liquidity policies. While important in practice, we abstract from this non-trivial extension in this paper and leave it for future research.

Fourth and finally, our benchmark framework studies a competitive banking sector while a deposit market power extension is introduced in Section 6.3. However, our model abstracts from imperfect competition in the *loan* market. Corbae and D'Erasmo (2021) (CD, henceforth) study banking industry dynamics in an empirically consistent, quantitative macro-banking environment with loan market power and credit markups. CD also compute a competitive equilibrium version of their model in order to compare to the benchmark with loan market power. Specifically, since CD's dominant-fringe model nests the perfectly competitive version as a special case, they set the cost of big bank entry large enough and calibrate the competitive model to match the same moments as in the benchmark.

3 Taking the Model to the Data

This section describes how we take our model to the data. One model period corresponds to a quarter. Table 1 summarizes the values of all parameters. Table 2 reports calibration targets and select moments in the model and in the data.[6] See Appendices A.1 and A.2 for more details on the data and the calibration targets.

3.1 Parameterization

We begin with the parameters that pertain to household preferences. We set relative risk aversion, ψ, as well as the Frisch labor supply elasticity, $\frac{1}{\chi_2}$, to 1 following the literature (Kaplan et al., 2018). The discount factor β is set to 0.996 in order to target an annualized

[6]While the discussion focuses on the baseline economy with ex-ante and ex-post bank heterogeneity, we note that every relevant variation of the model is separately re-calibrated in order to hit the same target moments.

Table 1: Model Parametrization

Parameter	Value	Description	Target/Source
		Households	
ψ	1	Risk aversion	Literature
β	0.996	Discount factor	Interest rate
χ_1	18.6	Labor disutility	Average hours 0.3
χ_2	1	Labor supply elasticity	Literature
		Firms and Technology	
α	0.33	Capital share	Literature
a	4.59	Production technology	Capital price 1
b	0.75	Production technology	Price elasticity of lending 0.25
		Banks	
σ	0.973	Bank survival rate	Gertler and Kiyotaki (2010)
λ	0.1295	Share of divertible assets	Average leverage 6.5
ζ_1	3.00E-05	Non-interest expsense, linear	Non-interest cost to assets ratio 0.015
ζ_2	2	Non-interest expsense, quadratic	Normalization
α_κ	1	Permanent types, Pareto	Zipf's law in Call Reports data
		Stochastic Processes	
σ_ξ	0.085	Volatility, idiosyncratic risk	Call Reports data
ρ_ξ	0.553	Persistence, idiosyncratic risk	Call Reports data
$\{A_L, A_H\}$	$\{0.994, 1.006\}$	States, aggregate risk	Output volatility 0.01
$\{\pi_{LL}, \pi_{HH}\}$	$\{0.9, 0.9\}$	Transition matrix, aggregate risk	Literature
		Additional Parameters for Extensions	
μ_H	0	Countercyclical return mean, high state	Call Reports data
μ_L	-0.02	Countercyclical return mean, low state	Call Reports data
λ_H	0	Countercyclical returns skewness, high state	Call Reports data
λ_L	-0.5	Countercyclical returns skewness, low state	Call Reports data
ν_b	0.12	Deposits in utility	Average markdown 0.8
θ_b	38	Elasticity of substitution across deposits	Markdown-size elasticity -0.0067

Notes: Model parameter values and brief descriptions.

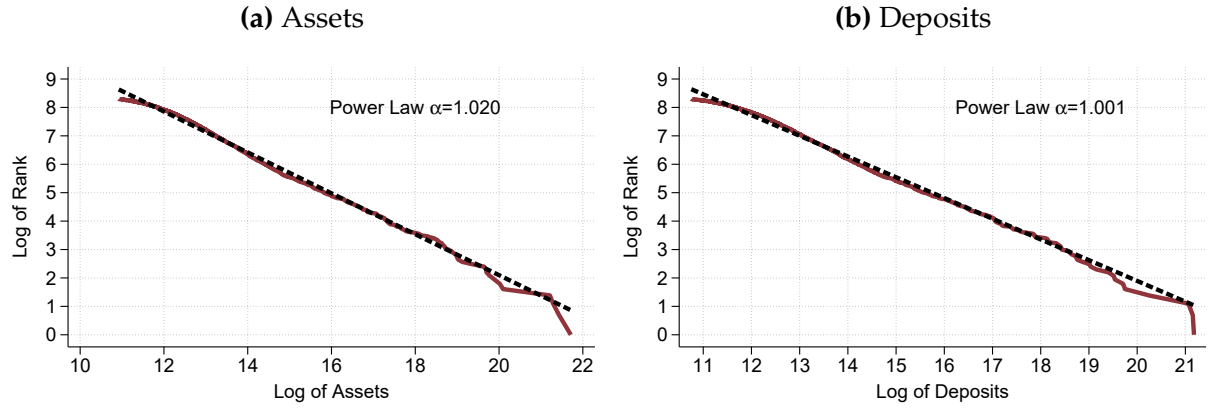
risk-free rate of roughly 1.6%. The labor weight χ_1 is internally calibrated so as to target average hours worked of one third.

On the side of technology and firms, we assume that the elasticity of output to capital is $\alpha = 0.33$, which is standard. The capital production function is assumed to take the following form:

$$\Phi(L) = a(L)^b \quad a, b > 0 \tag{25}$$

We calibrate the parameter a internally such that the average price of aggregate capital Q_t equals 1. Parameter b is set to 0.75 so that the elasticity of the price of capital to bank lending is equal to 0.25. This value corresponds to the elasticity of the price of capital with respect to firm investment that is typically found in the literature (Gilchrist and

Figure 1: Granular Banks in the Data



Notes: Panels (a) and (b) plot log-rank and log-size (solid lines), along with lines of best linear fit (dashed lines), for assets and deposits, respectively. The sample is restricted to the 4,000 largest U.S. banks as of 2020:q1.

Himmelberg, 1995; Gertler et al., 2020).

We now turn to the banking block. The bank survival rate σ is set to 0.973, following the literature (Gertler and Kiyotaki, 2010). The share of divertible assets λ is internally calibrated in order to hit an average leverage ratio—defined as claims over net worth in the model and total loans over equity in the data—of 6.5. This value is obtained from the population of commercial banks in the U.S. Call Reports. The linear parameter of non-interest expenses, ζ_1, is calibrated internally in order to hit the average non-interest costs to total loans ratio of 1.5%. The target is obtained from the same Call Reports sample of commercial banks. The power parameter, ζ_2, is chosen such that the cost function is quadratic.

Parameter α_κ is key for our analysis as it governs the shape of the distribution of permanent bank rate of return types κ. We proxy α_κ directly from the Call Reports data. Consider $\Pr(\text{Size} > S) = k/S^{\alpha_\kappa}$, which means that the likelihood of a bank being greater than some S is proportional to $\frac{1}{S^{\alpha_\kappa}}$. We estimate α_κ with maximum likelihood methods for both total assets and total deposits, for 2020:q1, and for the largest 4,000 banks, and obtain the values of 1.02 and 1.001 for assets and deposits, respectively.[7] Additionally, consider a rank-size rule that compares log-size against log-rank (Gabaix, 2009). A special case of this comparison is the Zipf's law, which arises if the relationship is approximated with a straight negatively-sloped line. We run the rank-size test for the U.S. banking sector based on the same sample as above.

Figure 1 plots log-rank of assets (panel A) and deposits (panel B) on the y-axes against log-size on the x-axes. It also reports the estimated power law parameters $\hat{\alpha}_\kappa$. The striking

[7] Results are highly robust across time.

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. Thus, based on these two pieces of evidence, we conclude that the Zipf's law is a good representation of the U.S. banking data and set α_κ to 1 in our model. This result is consistent with the existing estimates in the literature (Janicki and Prescott, 2004).

In practice, we discretize the distribution of types with eleven points, which is loosely consistent with a deciles interpretation. We draw a large number of observations from a Pareto density with $\alpha_\kappa = 1$ and compute eleven percentiles of the resulting draw. These percentiles correspond to the eleven types of $\kappa(j)$ that we use in the model. The draw is normalized through the constant k_m such that the median (sixth) type has a $\kappa(j)$ of unity. See Appendix C for further details on the numerical procedures.

We next turn to the parameterization of stochastic processes. To calibrate the persistence, ρ_ξ, and volatility, σ_ξ, of transitory return risk, we bring our return process (11) 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 residualize (log) RoL from the time fixed effect and run a linear regression on a bank fixed effect and an AR(1) component in the error term. We estimate a $\hat{\rho}_\xi$ of 0.553 and $\hat{\sigma}_\xi$ of 0.085. These are the values that we assign to the model parameters.[8]

We calibrate aggregate uncertainty in the model as follows. First, the transition probability matrix is set to $\boldsymbol{\pi}_a = \begin{Bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{Bmatrix}$. These values are chosen parsimoniously such that durations of good and bad times are the same. Second, we calibrate A_L and A_H internally in order to target a standard deviation of aggregate output, Y_t, in the recursive competitive equilibrium, of 0.01. This value corresponds to the volatility of U.S. GDP fluctuations and is standard in the literature.

Finally, the bottom panel of Table 1 reports parameters that are relevant for the two extensions of the baseline model analyzed below: countercyclical bank return risk and market power in the deposit market. We skip the discussion of these parameters for now and return to them in Sections 6.1 and 6.3, respectively.

3.2 Targeted and Untargeted Moments

Table 2 summarizes our calibration targets and reports select moments of the model and data. Our calibration procedure is able to hit all four of our targets with great precision.

[8]The empirical banking literature typically finds that uninsured idiosyncratic bank return risk is highly non-persistent (Galaasen et al., 2021).

Table 2: Calibration Targets and Select Moments

Moment	Model	Target / Data
Hours worked (target)	0.3	0.3
Average non-interest cost to loans ratio (target)	0.016	0.015
Average bank leverage (target)	6.5	6.5
Bank Assets Gini	0.69	0.94
Bank Deposits Gini	0.69	0.93
Marginal Propensity to Lend	6.26	
σ_Y (target)	0.01	0.01
σ_L/σ_Y	1.27	2.46
σ_N/σ_Y	1.43	1.90
σ_{Lev}/σ_Y	1.68	3.03
σ_{MPL}/σ_Y	10.37	
$\rho_{L,Y}$	0.97	0.36
$\rho_{N,Y}$	0.56	0.10
$\rho_{Lev,Y}$	0.90	0.23
$\rho_{MPL,Y}$	0.89	

Notes: Targets for calibration and select moments in the baseline model and the data. σ_x and $\rho_{x,y}$ stand for standard deviation and pairwise correlation coefficients, respectively. Banking moments are computed from U.S. Call Reports. Aggregate moments are computed from St. Louis Fed economic data.

First, the average number of hours worked, the average non-interest costs to loans ratio, and the average leverage ratio are all on target. Second, output volatility is 0.01 in the model as it is in the data.

The Table also reports several untargeted moments of interest. First, we report Gini coefficients for the model-implied distributions of bank assets and deposits. In the model, we obtain a Gini coefficient of roughly 0.7. The model generates a realistic, concentrated distribution of bank size. In the banking data, however, the Gini coefficients are above 0.9 as of 2020:q1 and in most years. Despite a considerable, order-of-magnitude improvement over the representative-bank benchmark, our baseline economy cannot fully account for the extreme levels of concentration in the U.S. banking sector. For example, it is 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. One prominent approach to further improve the distributional fit is to introduce non-measure-zero banks. Corbae and D'Erasmo (2023) develop a "quality ladder" model with banks that are not of measure zero, which subsequently widens the right tail of the size distribution. In addition, a feature of the data that could be added to 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.

Second, Table 2 shows that the average marginal propensity to lend (MPL) in our model is 6.26. This means that in response to a 1% positive shock to bank net worth, the average bank will increase lending by 6.26%. The magnitude is intuitive and corresponds closely to the average leverage ratio, which is 6.5. Interestingly, we have not found any empirical estimates of the MPL. Note that the MPL object captures bank-level lending responses to *bank-level* net worth changes. The empirical banking literature has plenty of estimates of bank-level lending responses to *aggregate* shocks. Estimating bank-level MPLs remains therefore a fruitful avenue for future research.

Third, Table 2 documents several untargeted business-cycle moments. The model-implied volatilities of aggregate claims, net worth, and bank leverage are all in the empirical ballpark. Importantly, the banking sector in the model is considerably more volatile than aggregate output, which is always the case in U.S. data. Time-series correlations of aggregate lending, net worth, and leverage vis-a-vis aggregate output are also close to but on average larger than in the data. Importantly, in our model both bank assets and leverage are considerably more pro-cyclical than bank net worth, which is also always true in U.S. data.

Finally, the Table also reports volatility and correlation with output for the average MPL. The former is around 10% while the latter is around 0.9. In other words, the aggregate model-implied MPL is highly contemporaneously procyclical and volatile. Again, as discussed above, we have not found any counterparts for these two moments in the empirical literature and hope that future research can fill this gap.

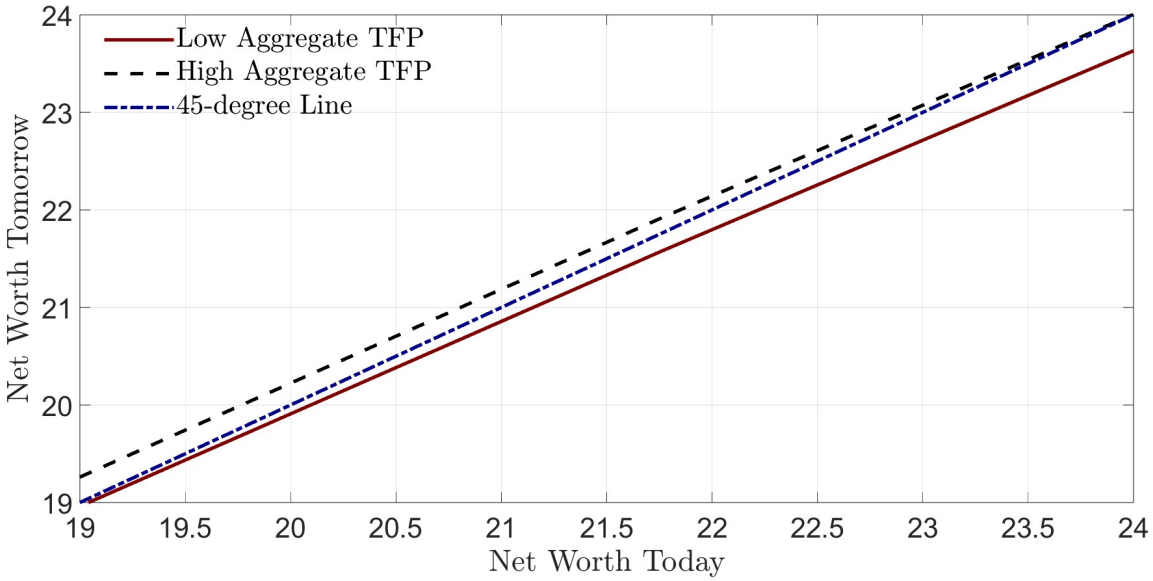
4 Model Properties

In this section we study selected properties of the model that are critical for the understanding of core mechanisms. First, we present and discuss aggregate and bank-level policy functions. Second, we discuss the marginal propensity to lend and its role as a sufficient statistic for aggregate fluctuations. Third and finally, we further elaborate on the banking distribution and its concentration.

4.1 Aggregate and Bank-level Laws of Motion

We begin with a key property of our model—approximate aggregation, whereby the dynamics of the banking distribution can be approximated with a small number of moments. Figure 2 displays future aggregate net worth against current aggregate net worth as implied by the recursive equilibrium of the model with ex-ante and ex-post bank het-

Figure 2: Tomorrow's vs Today's Aggregate Net Worth

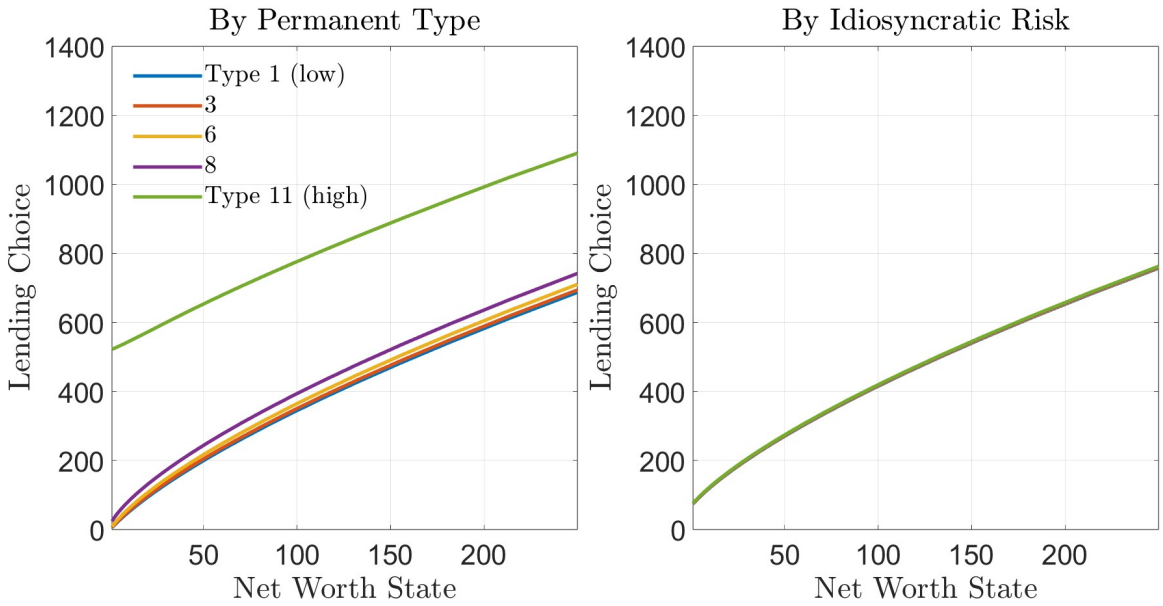


Notes: Tomorrow's vs. today's aggregate bank net worth for the baseline model. The top (bottom) line corresponds to the law of motion of aggregate net worth in a high (low) TFP state. The middle line is the 45-degree line.

erogeneity. The top and bottom lines in the Figure represent laws of motion for aggregate net worth in good and bad aggregate states, respectively. The middle line is the 45-degree line. Clearly, the aggregate net worth policy function is linear and the fit is very good. This is reminiscent of the same property characterizing laws of motion for aggregate capital in the model with heterogeneous households and incomplete insurance of Krusell and Smith (1998), KS henceforth. That property is at the root of the "approximate aggregation" result of KS whereby agents make very small mistakes when using the simple log-linear forecasting rule with just one moment of the distribution, such as in (18). Remarkably, approximate aggregation seems to also hold even in much richer environments such as Krueger et al. (2016) who extend the KS economy with ex-ante household heterogeneity and counter-cyclical earnings. In Appendix C.2 we elaborate upon this point further and provide formal measures of fit—the model R^2—and run other accuracy tests. Overall, this insight validates our approach of tracking the first moment of the distribution of bank net worth.

Figure 3 describes how the bank policy function for assets, $\mathcal{L}(n, \kappa, \xi)$, behaves across states in the equilibrium. We present two cases to illustrate the mechanism. In both panels, aggregate net worth is given at 21.76 and the aggregate state is good. The left panel shows the lending choice as a function of n and κ for the average value of ξ, while

Figure 3: Bank-level Lending Policy Function

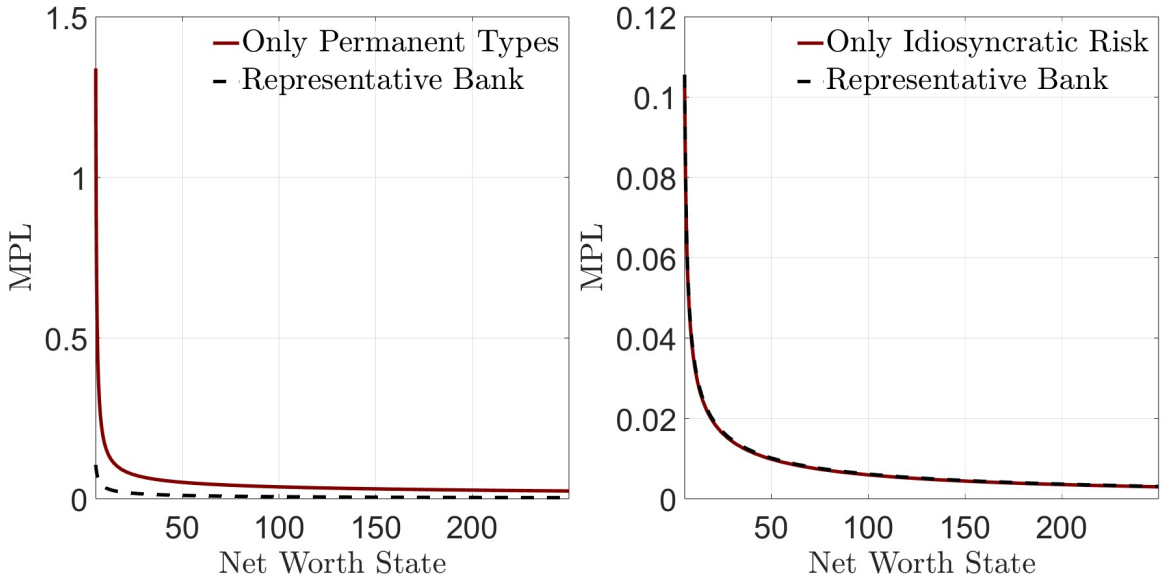


Notes: An individual bank's lending decision rule, $\mathcal{L}(n, \kappa, \xi; \bar{N}, A)$, for the baseline model with ex-ante and ex-post bank heterogeneity. Aggregate net worth, $\bar{N}$, is set at 21.76. Aggregate productivity is in the good state. The left panel averages out idiosyncratic risk, ξ. The right panel averages out permanent heterogeneity, κ.

the right panel shows the lending choice as a function of n and ξ for the average value of κ.

Three aspects are worth noticing. First, in all cases lending is an increasing function of net worth. Second, for a given value of net worth, lending is increasing in the permanent type, κ. Because κ is Pareto-distributed, banks that belong to the highest, 11th, type are disproportionally more efficient and are capable of lending a lot more for the same unit of net worth. Third, as is clearly seen from the right panel, idiosyncratic risk, ξ, is approximately irrelevant for the lending choice. This is strongly reminiscent of a canonical "as-if" result (Krusell and Smith, 1998; Werning, 2015): in the presence of *only* incomplete markets and uninsured idiosyncratic return risk, the economy behaves as if it is populated by a single bank. The reason for this result is that, even though, in the presence of incomplete markets and idiosyncratic shocks, the banking distribution still moves around, most of net worth is accumulated by banks with the same slope of the lending curve, yielding exact aggregation.

Figure 4: Marginal Propensity to Lend



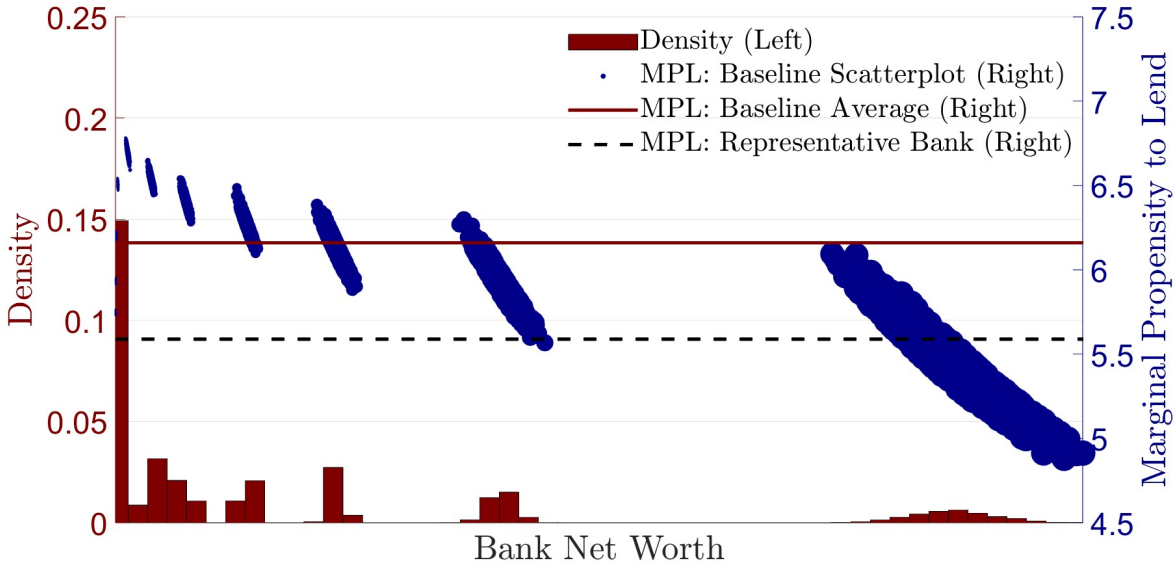
Notes: Marginal propensity to lend (MPL) functions in various models (y-axis), as a function of bank-level net worth (x-axis). The left and right panels plot MPL functions for the models with only permanent heterogeneity, κ, and idiosyncratic risk, ξ, respectively. In the respective cases, the κ and ξ dimensions have been averaged out. Dashed lines correspond to the MPL function in the representative-bank economy.

4.2 Marginal Propensity to Lend

Figure 4 displays the marginal propensity to lend (MPL) as a function of bank net worth. Recall that the MPL, as shown in equation 17, is defined as the change in bank-level lending in response to a marginal change in bank-level net worth. In the left panel, we solve the special case of our model with just permanent heterogeneity and without idiosyncratic shocks. In the right panel, conversely, we consider the case without permanent heterogeneity and with idiosyncratic shocks.

We observe that in both situations the MPL is decreasing in bank net worth: smaller banks have a greater elasticity of lending with respect to shocks to net worth. This property is consistent with a classic evidence on the heterogeneous effects of the bank lending channel (Kashyap and Stein, 1995, 2000). Crucially, for any level of net worth, the MPL in the economy with permanent heterogeneity is greater than the MPL in a representative-bank benchmark (left panel). Noticeably, in the version of the model with ex-post heterogeneity only (right panel) the MPL is approximately identical to the one in a representative-bank model. This result is related to the earlier discussion on the lending policy function and re-confirms the as-if property: idiosyncratic shocks by themselves

Figure 5: Marginal Propensity to Lend in the Distribution



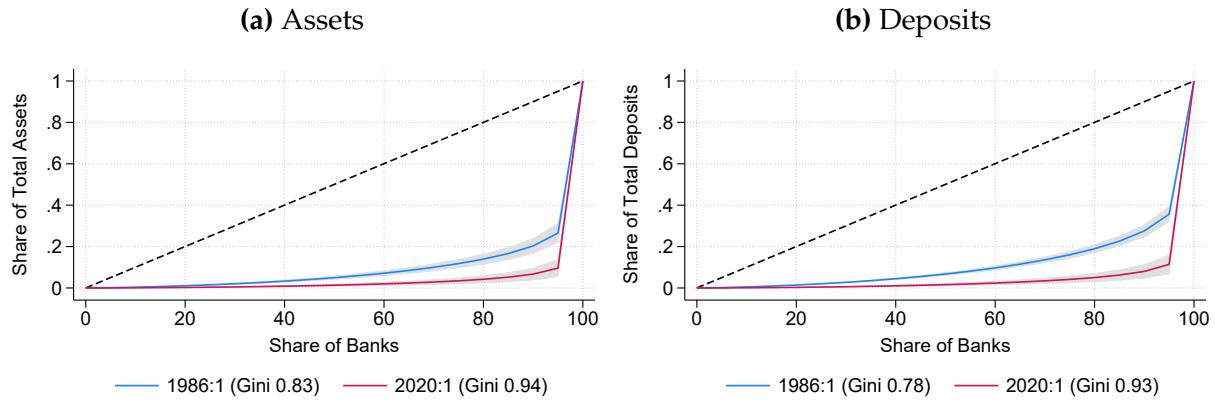
Notes: Marginal propensity to lend (MPL) across the distribution of bank-level net worth in the recursive equilibrium of the baseline model with both ex-ante and ex-post heterogeneity. Straight and dashed horizontal lines refer to average MPLs in the baseline economy and the representative-bank benchmark, respectively. The left y-axis shows the density of the distribution and the right y-axis shows the MPLs in percentage points.

deliver an aggregate lending elasticity that is hardly distinguishable from that of a model with just a representative bank.

We now illustrate how the MPL behaves across the distribution of banks in the recursive equilibrium of the model with both ex-ante and ex-post heterogeneity. Figure 5 plots four objects: the density of bank net worth, the MPL of every bank in the baseline economy, and the average MPLs of the baseline and the representative-bank models.[9] Three results are worth emphasizing. First, the net worth density has a power law shape, with a high share of small (low net worth) banks and a low share of very large banks (high net worth). We return to this relevant property of banking inequality and concentration in the next section. Second, as already alluded to before, the MPL is decreasing in bank net worth. Third, the average MPL in the baseline economy is larger than the MPL in a corresponding representative-bank economy. Heterogeneity introduces a mass of banks that are very large and with a low MPL. However, their share is not great enough to counteract the larger mass of small, high-MPL banks. As a result, the average MPL in our baseline economy is larger relative to the representative-bank counterpart.

[9]The shown histogram of net worth is for a given period in the simulation. Figure B7 in the Appendix presents a waterfall graph that shows the evolution of the distribution over time.

Figure 6: Banking Concentration in the Data



Notes: Panels (a) and (b) plot Lorenz curves and report Gini coefficients for total assets and total deposits in select years, respectively. Data is from the U.S. Call Reports. The sample includes commercial banks.

In anticipation of the next sections, the MPL heterogeneity channel is the key reason why business-cycle fluctuations—and particularly the response of total bank lending—are amplified in our economy relative to the representative-bank benchmark.[10]

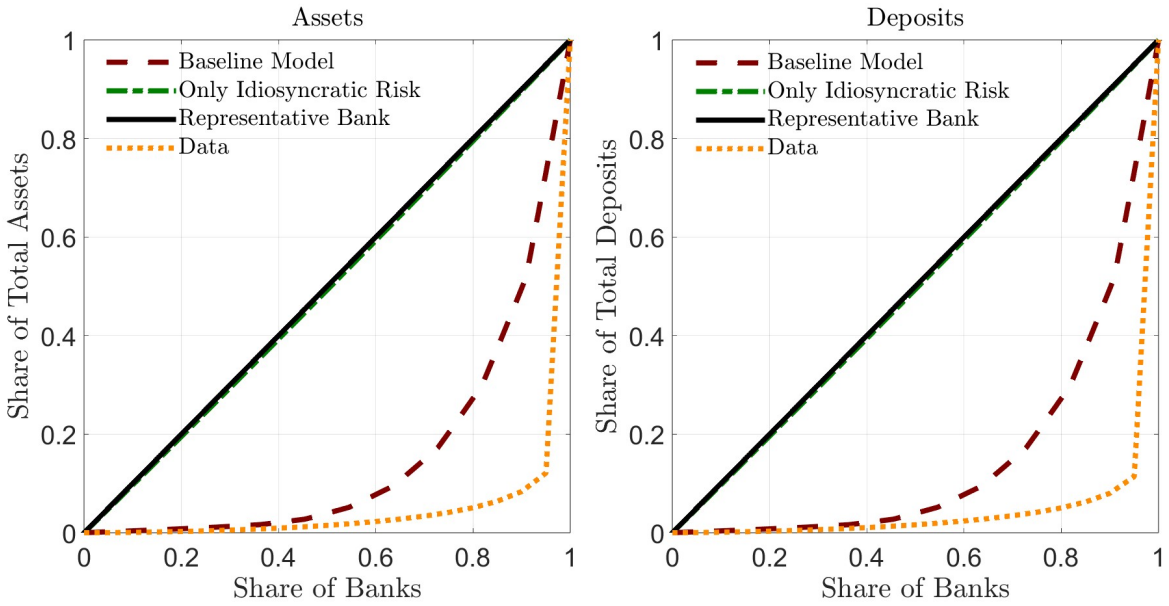
4.3 Bank Size Distribution and Concentration

A robust feature of U.S. banking data is the high level of market concentration. It is not only considerably high at present times but has been rising since at least the 1980s (Corbae and D'Erasmo, 2020). Figure 6 plots Lorenz curves—a standard market concentration metric—for commercial bank total assets, using Call Reports data. The 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 control 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. In panel (b), Figure 7 plots the Lorenz curves for the U.S. commercial bank (domestic) deposit market, also using data from the Call Reports. The rise of banking concentration is even more significant when size is

[10]Conceptually, the MPL object is related to several constructs in the literature. For example, the marginal propensity to consume (Kaplan and Violante, 2022), the marginal propensity to take on risk in an environment where heterogeneous households choose their portfolio risk exposure (Kekre and Lenel, 2022), or the marginal propensity to invest in models with heterogeneous firms and financial frictions (Ottonello and Winberry, 2020).

Figure 7: Banking Concentration in the Model



Notes: Panels (a) and (b) plot Lorenz curves for bank claims, l, and deposits, b, in various versions of the model, respectively. Dotted lines correspond to U.S. Call Reports data, 2010:q1.

proxied with deposits. Consider that the deposits Gini coefficient has grown by about 20% since 1986 and is currently at 0.93. Generally speaking, choosing the right 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 (Drechsler et al., 2017).

Figure 7 visualizes our model's ability to generate a realistic degree of size concentration. It plots the model-implied Lorenz curves for total assets (left panel) and total deposits (right panel) in four different cases: (i) the baseline model with ex-ante and ex-post heterogeneity, (ii) the benchmark case with a representative bank, (iii) the case with only ex-post heterogeneity, and (iv) U.S. data, corresponding to 2020:q1. Any departure from the perfect equality counterfactual (the representative-bank case) suggests that the banking sector features some degree of inequality. Notice that the baseline economy is considerably concentrated with Gini coefficients of around 0.70 for both assets and deposits; conversely the economy with just idiosyncratic risk displays a negligible degree of concentration. The representative-bank counterfactual, in both cases, features a Gini coefficient of exactly 0, by construction.

In the model with just ex-post heterogeneity, the bank net worth distribution has properties at odds with the data: there are too few low net-worth banks and there is too limited

concentration among the very high net-worth banks. Hence, net worth is accumulated by banks with approximately the same slope of the lending curve, leading to approximately symmetric lending choices. With the introduction of permanent heterogeneity, however, this symmetry breaks as the slope of the lending curve varies across the distribution. This insight is at the heart of the previously discussed "as-if" result, whereby banks' choices are approximately symmetric.

In summary, in this section we have discussed four properties of our framework. First, given the linearity of the aggregate bank net worth law of motion, the model features approximate aggregation—the aggregate productivity state and the mean of the net worth distribution are sufficient statistics to describe aggregate dynamics. Second, permanent heterogeneity breaks the symmetry of bank-level lending decisions and MPLs, yielding a departure from the representative-bank benchmark. Third, MPLs are systematically heterogeneous: declining in size and increasing in portfolio rates of return. The average MPL of the model with heterogeneous banks is greater than the MPL of a representative bank. Fourth and finally, our model delivers a realistically concentrated distribution of bank size with a small number of banks controlling a disproportionally large share of assets, deposits, and net worth.

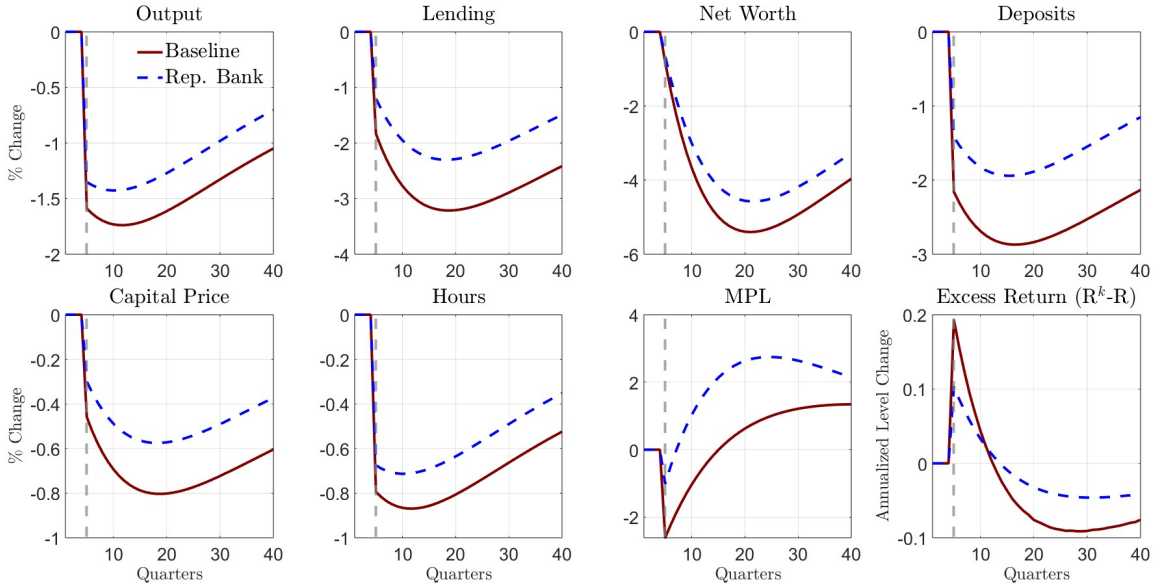
5 Aggregate Fluctuations

In this section, we study the model's dynamic responses to aggregate and idiosyncratic shocks. We begin by analyzing the model's behavior in reaction to aggregate Total Factor Productivity (TFP) shocks, followed by an examination of the underlying mechanisms to isolate the contributions of different channels. Next, we assess the model's response to granular shocks, specifically idiosyncratic disturbances affecting only a particular subset of banks. Finally, we conduct long-run simulations of the model to study the occurrence and impact of banking and economic crises.

5.1 Impulse Responses to TFP Shocks

We begin by comparing impulse responses to a negative aggregate TFP shock under two scenarios: the baseline economy with heterogeneous banks (featuring both ex-ante and ex-post heterogeneity), and the representative-bank special case. To obtain these impulse response functions, we perform the following computational steps. First, we run a simulation based on an already-solved economy with 2,000 banks for $\mathbf{T} = 2,000$ periods using both aggregate and idiosyncratic shocks. We discard a fraction of earlier

Figure 8: Impulse Responses to an Aggregate TFP Shock



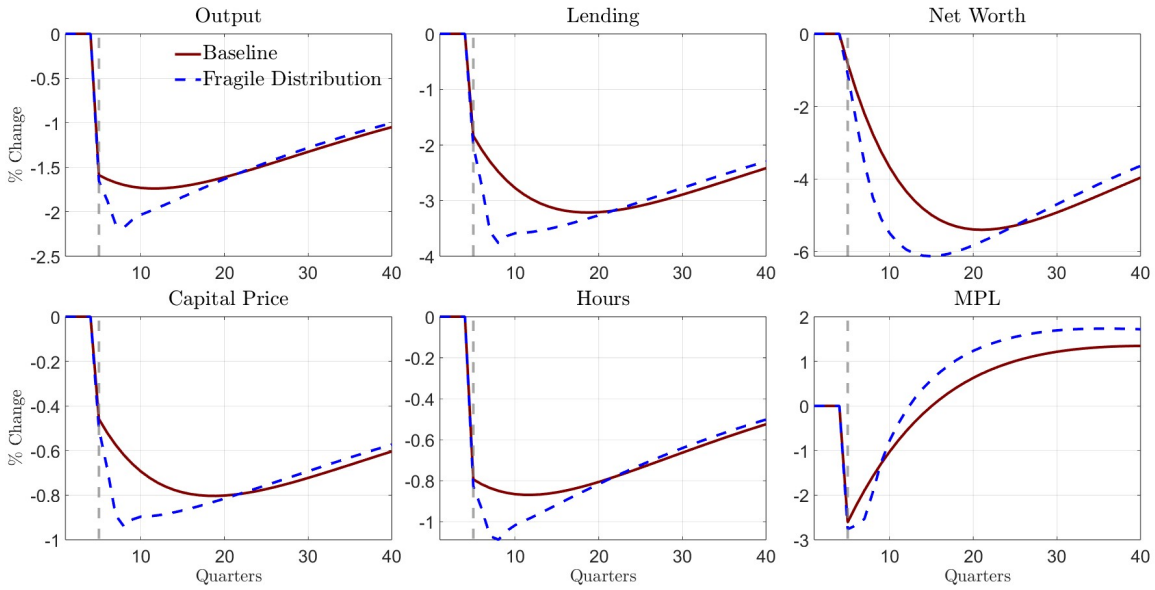
Notes: Impulse response functions to a one standard-deviation negative shock to aggregate TFP, A_t. The shock hits at $t = 5$. The straight and dashed lines correspond to the baseline economy with both ex-ante and ex-post bank heterogeneity and the representative-bank case, respectively.

periods. At time $\mathbf{T}^* < \mathbf{T}$—represented on the figures by vertical lines—there is a sudden one standard-deviation negative innovation to A_t that mean-reverts back to zero over time at a rate 0.95 (a "MIT shock"). Second, we run a second simulation which is identical to the first except that there is no MIT shock occurring at $\mathbf{T}^*$. Finally, to obtain impulse responses we subtract the path of aggregates in the second simulation from the path of aggregates in the first simulation.

Figure 8 presents the results. In the baseline economy, a negative TFP shock generates a severe financial tightening: the economy-wide credit spread (excess return on capital) spikes and the banks' franchise value falls. As a result, the size of the banking sector shrinks, with assets, deposits, and net worth all contracting. The aggregate marginal propensity to lend (MPL) falls on impact due to lower returns, but recovers quickly because of the persistent deterioration in bank net worth, which is negatively associated with the MPL in the distribution. The bank lending channel transmits onto non-financial firms which, upon receiving less funding from the banks, produce less capital, pushing the price of capital downward. As a result, an aggregate recession ensues: total output, consumption, and working hours all decline.

Relative to the behavior in the representative-bank case, the baseline economy exhibits

Figure 9: Financial Fragility and Distributional State Dependency



Notes: Impulse response functions to a one standard-deviation negative shock to aggregate TFP, A_t, conditional on a prior negative financial shock. The TFP shock hits at $t = 5$. The financial shock hits at $t = 4$ and lasts for three quarters. The straight and dashed lines correspond to the baseline economy with both ex-ante and ex-post bank heterogeneity and with and without the financial shock, respectively.

noticeable amplification of both financial and macroeconomic aggregates, particularly bank lending, deposits, and total output. As we can see, output contracts by 20 per cent more in the case with bank heterogeneity relative to the representative-bank limit. The intuition for this result is best understood by again recalling MPL heterogeneity. As Figures 4 and 5 have documented, the average MPL in the baseline economy with permanent and stochastic heterogeneity is higher than the MPL of the representative intermediary because of the presence of a large number of small and highly sensitive banks. This leads to a greater average ex-ante economy-wide MPL and a substantial ex-post decline in net worth, bank lending, and production.[11]

5.2 Financial Fragility

In our environment, the endogenous distribution of bank net worth, Γ_t, is a relevant state variable. As such, all aggregate responses to exogenous shocks depend explicitly on its

[11] 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; Auclert et al., 2024).

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 term the "fragile" economy. In the fragile economy, a negative financial shock in period $\mathbf{T}^* - 1$ causes a leftward shift in the distribution of bank net worth. We assume that this shock persists for three quarters. Impulse responses for the fragile economy are derived similarly to before. Specifically, we run two model simulations, incorporating both aggregate and idiosyncratic shocks, with the fragility-inducing financial shock occurring at $\mathbf{T}^* - 1$. In the first simulation, we introduce a negative TFP shock at $\mathbf{T}^*$, while in the second simulation this shock is absent. The difference in responses between the two simulations provides the identification of interest: the impact of aggregate TFP shocks when the banking sector is in a normal state versus a fragile state.[12]

Figure 9 plots the results. 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 contraction. The excess contraction scales with the duration and amplitude 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 smaller and closer to the zero bound on net worth. As a result, any subsequent negative aggregate shock has larger real effects on the economy.

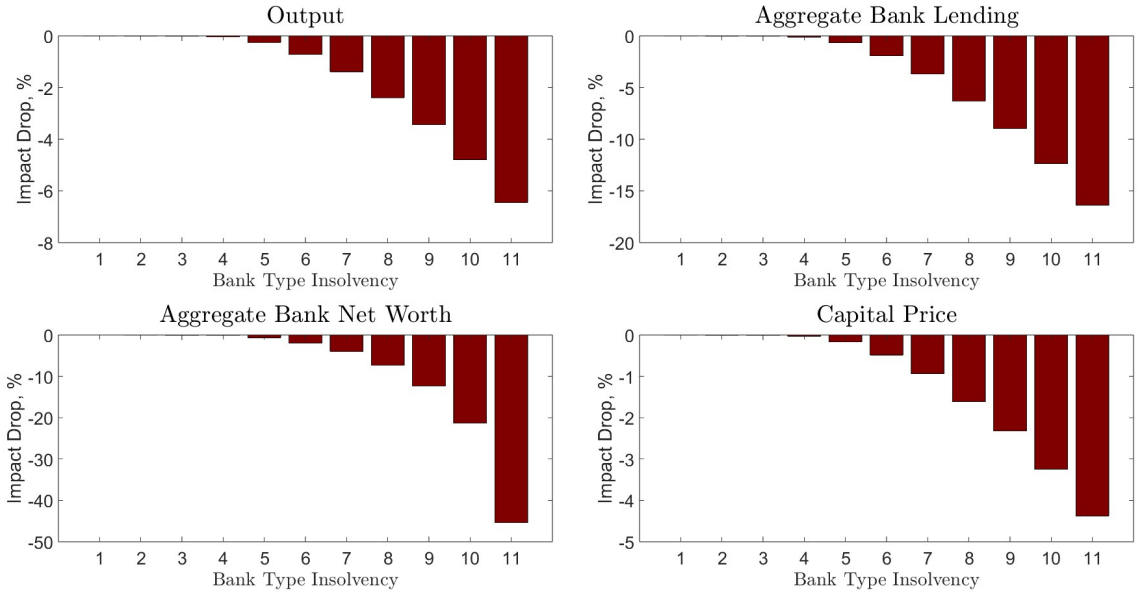
5.3 Granular Bank Failures

Recall that concentration is a significant feature of the banking sector—both in U.S. data and in our model, as illustrated in Figure 7. The seminal work by Gabaix (2011) introduced the "Granular Hypothesis," which posits that idiosyncratic firm-level disturbances can generate aggregate fluctuations, provided that some firms are abnormally large, or granular. This hypothesis has been extensively studied in the context of non-financial firms (Carvalho and Grassi, 2019), bank credit portfolios (Galaasen et al., 2021), and international trade (Gaubert and Itskhoki, 2021).

In practice, finance and banking regulators are concerned about the potential for a few large, granular banks to cause a macroeconomic crisis if they fail. This concern is often referred to as the "too-big-to-fail" (TBTF) hazard, suggesting that the collapse of certain large financial firms would have catastrophic consequences for the broader economy. The

[12]Numerically, we adjust the discrete grid in the direction of bank-specific net worth, n, by shifting it leftward by one point with the exception of the lower bound. This lasts for the duration of the financial shock.

Figure 10: Granular Bank Failures



Notes: The aggregate impact of failures of banks belonging to a specific permanent return type, κ. Failure corresponds to a shock that wipes out all net worth. The impact, in percentage points, is shown on the y-axes. Types, in order, are shown on the x-axes.

TBTF externality has been widely studied both empirically and theoretically (Farhi and Tirole, 2012; Philippon and Wang, 2023). However, many existing quantitative macro-banking models with a representative financial intermediary find it challenging to study this issue. Our framework, which incorporates heterogeneous banks, is well-suited to investigate this question.

To explore the TBTF problem, we conduct the following exercise. Using the equilibrium lending policy function, $\mathcal{L}(n, \kappa, \xi)$, we calculate the lending response of each type κ in the economy to a shock that reduces the net worth of all banks of that type to zero, while maintaining the equilibrium net worth of all other types. Since banks cannot operate with negative equity, this shock effectively represents bank failure.[13] From the lending response, we compute the impact on capital production, bank net worth in the subsequent period, the price of capital, and aggregate output.

Figure 10 presents the results. Consider the 11th (right-most) column in the top-right panel, which shows the lending response to the failure of all banks belonging to the 11th, most profitable type. The 11th column in the top-left panel shows the response of aggregate output to the same shock, whereas the bottom panels display the corresponding responses of bank net worth and capital price respectively. The results clearly demonstrate

[13]Importantly, as mentioned previously, we also assume that banks have a constraint on outside equity issuance which prevents them from replenishing net worth on-demand in the short run.

that the failure of granular banks has dramatic consequences for the economy. The higher the type, the larger the ex-ante net worth of these banks, and the more severe the macroeconomic effects resulting from their failure. The collapse of the highest-type banks triggers a significant real and financial crisis: output contracts by about 7 percent, lending drops by 17 percent, bank net worth plunges by around 45 percent, and the price of capital falls nearly 5 percent. In contrast, the failure of smaller, lower-type banks has negligible aggregate consequences.

We conclude that the TBTF hazard is indeed present in our model economy: the failure of granular banks can provoke severe crises and recessions, whereas the role of smaller banks in such outcomes is minimal.

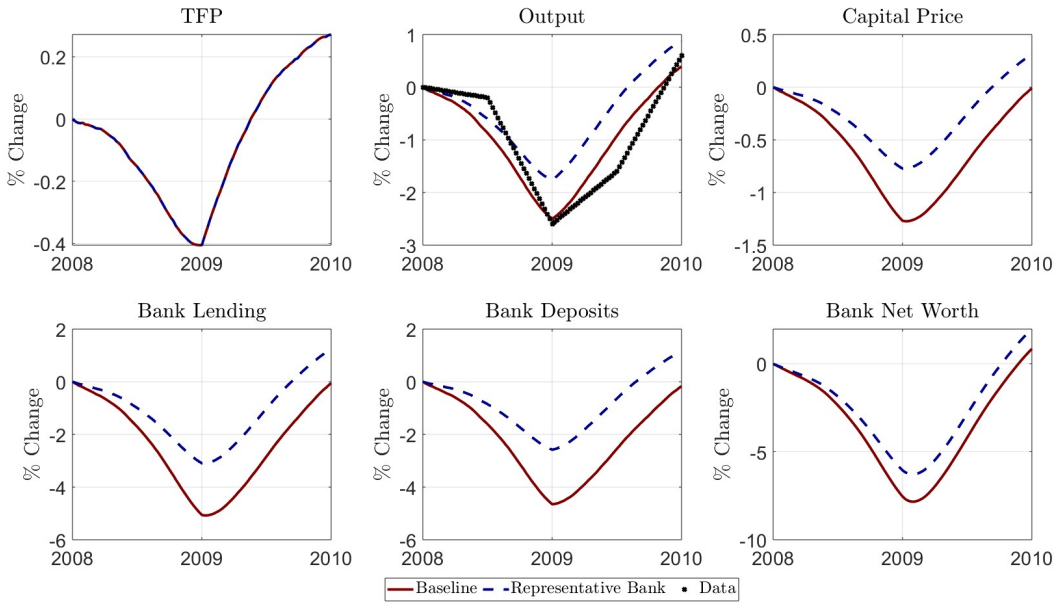
5.4 Financial Crises

In this section, we employ our framework to characterize banking and economic crises that may arise endogenously within a long simulation of the model. We adopt an event study approach similar to Calvo et al. (2006) and Mendoza (2010). Specifically, we simulate a sequence of aggregate and idiosyncratic shocks over 10,000 periods, generate a panel of 2,000 banks, and compute the aggregates using equilibrium policy and value functions in each period. We define a *crisis event* as a decline in aggregate output triggered by an aggregate Total Factor Productivity (TFP) shock falling below a specified threshold. All other paths are untargeted. For each event, we store the paths of all relevant variables, capturing both pre- and post-event periods. We then take averages across events, and perform the same calculations for a representative-bank benchmark.

In Figure 11, we compare the baseline economy with ex-ante and ex-post bank heterogeneity (solid line) to an economy with a representative bank (dashed line). Additionally, we plot the actual decline in U.S. real GDP from 2008-2010 for a quantitative comparison. The trough of 2009 is our empirical target, which we aim to match in the baseline model by selecting an appropriate shock threshold. Notably, the representative-bank economy underestimates the depth of the output decline, missing the trough by approximately one-third.

Our baseline model produces significantly sharper contractions in bank assets, deposits, and net worth during crisis episodes. Importantly, the decline in actual TFP is quantitatively similar across both cases. Therefore, the difference in the financial and economic aggregates' responses stems from internal amplification mechanisms within the model. This amplification arises directly from MPL heterogeneity, as discussed throughout the paper: the baseline model exhibits much greater sensitivity to aggregate shocks than the representative-bank case.

Figure 11: Simulating the Great Financial Crisis



Notes: Event study analysis of financial crisis events based on a long simulation of the baseline economy and of the representative-bank benchmark. An event is defined as a decline in aggregate output triggered by an aggregate Total Factor Productivity (TFP) shock falling below a specified threshold. The threshold is calibrated to match the 2009 empirical trough of output in the baseline model.

Overall, in this section we conducted a quantitative investigation of business-cycle dynamics in our model economy. We found that bank heterogeneity amplifies aggregate fluctuations, primarily through the MPL heterogeneity channel. Since the distribution of bank size (net worth) is an important state variable, our model exhibits aggregate distributional state-dependency: standard aggregate shocks are amplified when they hit an ex-ante fragile distribution. Finally, the high level of concentration of bank size gives rise to a "too-big-to-fail" property, as the failure of large banks can have considerable implications for the broader macroeconomy.

6 Model Extensions

In this section, we explore two major extensions of the baseline model. First, we incorporate a key empirical feature: the observation that idiosyncratic return risk faced by banks is counter-cyclical. Building on this extension, we also analyze granular economic cycles triggered by idiosyncratic shocks that specifically affect only large banks. Second, we relax the baseline assumption of perfect competition by introducing market power in the deposit market. In this variation, banks are allowed to exert pricing power over deposits. For both extensions, we provide empirical evidence that supports the relevance

and validity of these modifications.

6.1 Counter-Cyclical Return Risk

An extensive body of literature has shown that households and non-financial firms experience greater idiosyncratic risk during recessions, largely due to the pro-cyclicality of the third moment of income growth (Guvenen et al., 2014; Bloom et al., 2023). A natural question arises: do banks face a similar pattern of heightened idiosyncratic risk during economic downturns?

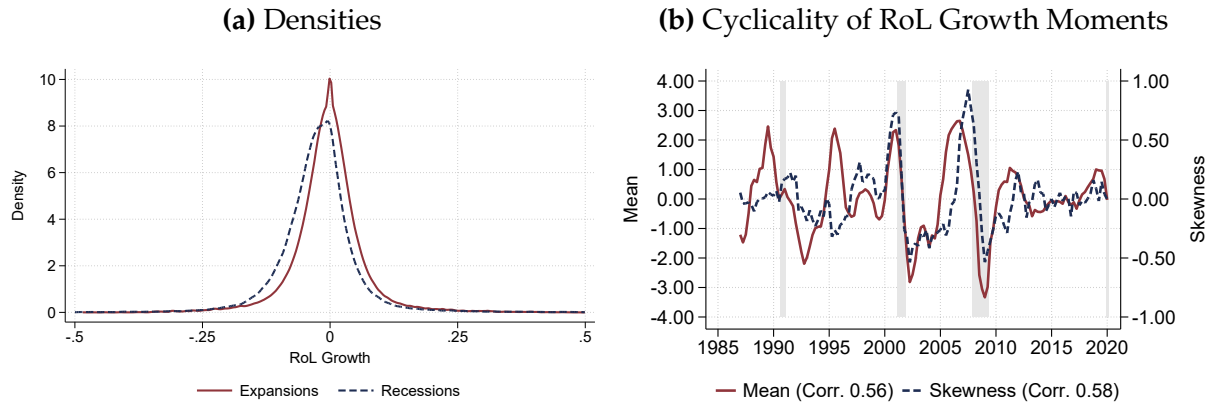
Counter-Cyclical Returns in the Data Recall that, in the baseline model, uninsured idiosyncratic bank rate of return risk, $\xi_t(j)$, follows an AR(1) process with shocks drawn from a Gaussian distribution. Now, we consider a situation where shocks $\xi_t(j)$ are possibly aggregate state-dependent and, in particular, counter-cyclical. To fix ideas, we first proceed with the data. We start with the same return on loans (RoL) measure that we compute from Call Reports data. The quarter-on-quarter log-difference of this variable, which we label $\Delta r_{j,t}$, constitutes our transitory loan return risk measure. This measure of loan income risk is built in line with the literature (Guvenen et al., 2014).

Figure 12 depicts the distributions of $\Delta r_{j,t}$ for U.S. expansions and recessions, defined by the National Bureau of Economic Research (NBER) criterion, and over the 1984:1-2020:1 period. The density of $\Delta r_{j,t}$ is visibly pro-cyclical. 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 moments. The third moment is defined by 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 of the resulting smoothed measures of the first and third moments. Both variables are strongly pro-cyclical, with pairwise correlations with U.S. real GDP growth equal to 0.56 and 0.58, respectively, and statistically significant at the 1% level. A pro-cyclical left skewness implies that the idiosyncratic rate of return risk faced by banks is *counter-cyclical*. Put differently, 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.[14]

The pro-cyclicality of the skewness of bank income risk is a novel and important finding. It suggests that banks face greater downside risk in recessions. The implication

[14]Interestingly, we find that the second moment of the distribution of RoL growth, defined by the standard deviation, is essentially flat over the business cycle. This finding is consistent with the literature (Busch et al., 2022).

Figure 12: Counter-Cyclical Returns in the Data



Notes: Panel (a) plots densities of log-differenced bank returns on loans, $\Delta r_{j,t}$, for U.S. expansions and recessions separately. Panel (b) plots the time series of the first and third moments of $\Delta r_{j,t}$; both series have been HP-filtered and smoothed with a moving-average filter with four lags.

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. This finding is consistent with and complements a growing literature that points to the importance of the *third* moment in the dynamics of earnings growth for households and firms (Guvenen et al., 2014; Bloom et al., 2023). All in all, the pro-cyclicality of the first and third moments of the banks' RoL growth distribution are key motivating facts of an extension of our baseline theoretical framework.[15]

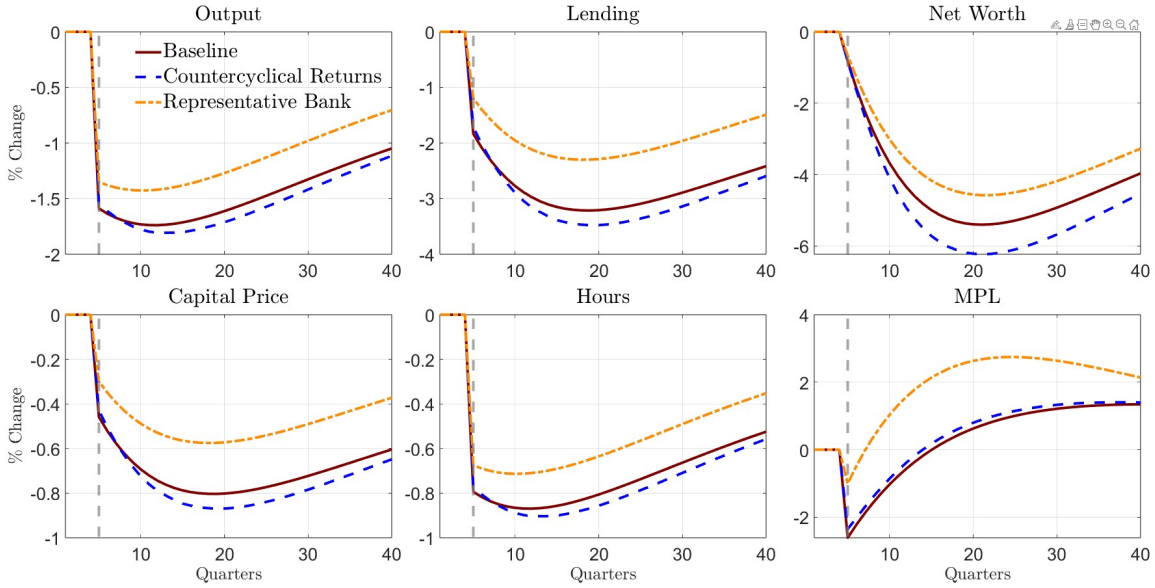
Counter-Cyclical Returns in the Model We now assume that the transitory rate of return risk, $\xi_t(j) \in \Xi$, follows an AR(1) process with shocks $\epsilon_t(j)$ drawn from a *non-Gaussia*n, aggregate state-dependent distribution. Specifically, we employ the Hansen (1994) Skewed-t density with time-varying 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:[16]

$$\xi_t(j) = (1 - \rho_\xi)\mu_\xi + \rho_\xi \xi_{t-1}(j) + \epsilon_t(j) \tag{26}$$

[15] Appendix A.4 provides several robustness tests for this empirical finding. First, we show that this result also holds at the bank *holding* level of aggregation. Second, in order to account for a possible influence of outliers, we plot the median and Kelley skewness. Finally, to show that our result is not driven by the filtering approaches, we also show the raw, unfiltered series of the first and third moments. Our conclusions do not change.

[16] Our approach complements other existing methods in the literature. For instance, see Bloom et al. (2023) who model $\xi_t(j)$ as a random variable that is drawn from a mixture of several normally distributed random variables.

Figure 13: Counter-Cyclical Returns in the Model



Notes: Impulse response functions to a one-standard deviation negative aggregate TFP shock in the baseline economy with ex-ante and ex-post bank heterogeneity and a-cyclical income, in the extension with counter-cyclical returns, and in the representative-bank benchmark.

and

$$\epsilon_t(j) \sim g(z_t(j)|\eta_t, \lambda_{\epsilon,t}) = \begin{cases} bc\left(1 + \frac{1}{\eta_t - 2}\left(\frac{bz_t(j)+a}{1-\lambda_{\epsilon,t}}\right)^2\right)^{-(\eta_t+1)/2}, & z_t(j) < -a/b \\ bc\left(1 + \frac{1}{\eta_t - 2}\left(\frac{bz_t(j)+a}{1+\lambda_{\epsilon,t}}\right)^2\right)^{-(\eta_t+1)/2}, & z_t(j) \geq -a/b \end{cases} \tag{27}$$

where $z_t(j)$ is a variable drawn from the standard Normal distribution with mean $\mu_{\epsilon,t}$, and the triad $\{a, b, c\}$ are 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: $\eta_t = L + \frac{U-L}{1+\exp(-\lambda_{\epsilon,t})}$ (where U and L are constants) such that the only source of time-varying deviation from normality is the skewness parameter $\lambda_{\epsilon,t}$. Figure B6 in the Appendix 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 $\lambda_{\epsilon,t} < 0$.

There are four new parameters that require calibration, as summarized in Table 1. First, μ_H and μ_L represent the means of the process z during good and bad aggregate states, respectively. We set these values at 0 and -0.02, reflecting the average levels of aggregate RoL growth during U.S. expansions and recessions, as illustrated in Figure 12. Second, λ_H and λ_L correspond to the skewness parameter, $\lambda_{\epsilon,t}$, for good and bad aggregate states. We assign values of 0 for λ_H and -0.5 for λ_L, based on the observation in Figure 12 that the skewness of RoL growth decreases by about 0.5 points during recessions. Therefore, we

assume that idiosyncratic shocks ξ are drawn from a Normal distribution in good states and a left-skewed distribution in bad states.

With the Skew-t process calibrated, we can now analyze aggregate dynamics in the model with counter-cyclical return risk. We consider impulse response functions to a negative aggregate TFP shock under two scenarios. The first scenario is the baseline model with Gaussian, a-cyclical income risk. In the second scenario, when the negative TFP shock hits at $\mathbf{T}^*$, both the mean and skewness of the idiosyncratic shock distribution shift from μ_H and λ_H to μ_L and λ_L, respectively. Both parameters gradually revert to normal levels with a mean-reversion rate of 0.95. During this period, all idiosyncratic shocks ξ are drawn from the left-skewed distribution.

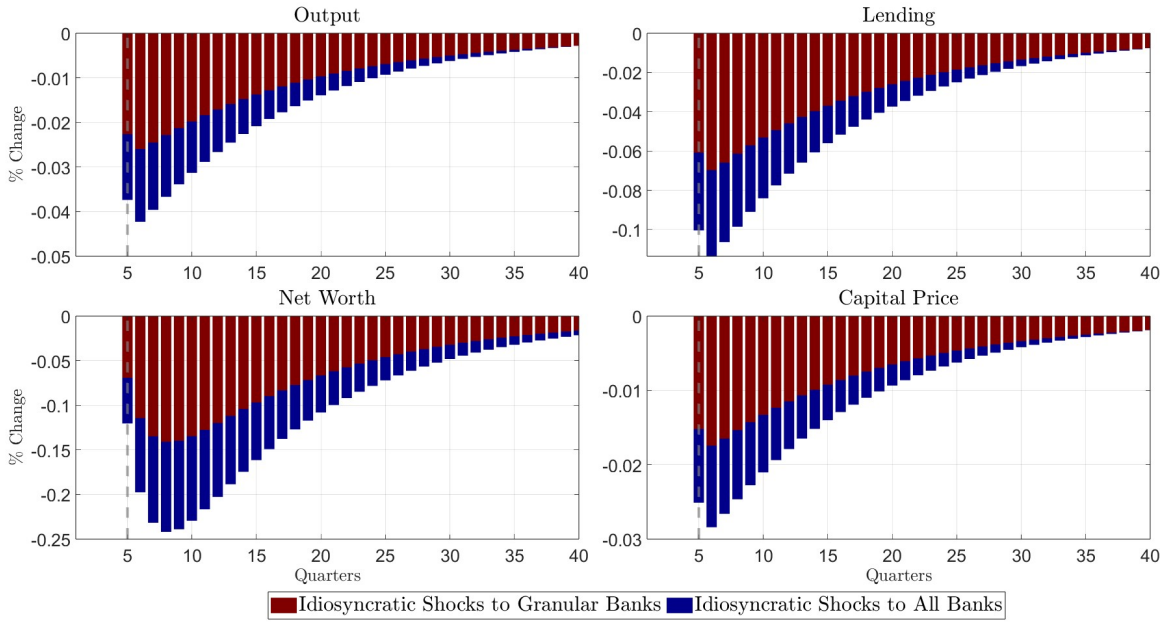
Figure 13 shows the results. Counter-cyclical idiosyncratic bank returns amplify aggregate fluctuations compared to the baseline model with a-cyclical risk. The underlying intuition is that entering a recession shifts the distribution of idiosyncratic returns towards a more left-skewed shape, exposing banks to greater downside risk. As this downside risk materializes, a fraction of banks faces particularly poor portfolio outcomes, leading to contractions in bank lending, production slowdowns, and declines in hours worked. For comparison, the figure also presents impulse responses for the representative-bank benchmark. Compared to the standard model, our extended framework with counter-cyclical bank income risk exhibits significant amplification of aggregate shocks.

6.2 Granular Banking Cycles

We have previously demonstrated that bank heterogeneity can amplify aggregate fluctuations. But can *idiosyncratic* shocks exclusively to granular banks—those of the highest return type κ—generate business cycles? In other words, we are now explicitly testing whether our model can produce granular banking dynamics. To this end, we will be employing the version of the model with counter-cyclical transitory income risk introduced in Section 6.1. This exercise is complementary to, but distinct from, our earlier analysis of banking failures in Section 5.3.

In Figure 14, we present results of an exercise comparing impulse responses across two specifications. First, we simulate the model as before but introduce the following MIT shock at period $\mathbf{T}^*$: banks belonging to the 11th permanent return type, κ, experience a one-quarter reduction in both the mean and skewness of the idiosyncratic shock distribution, while aggregate TFP remains unchanged and no other agents in the economy are affected. We refer to this as a "granular shock." Second, we run another simulation in which all banks experience the same drop in the mean and skewness of their idiosyncratic shock distribution at $\mathbf{T}^*$. Finally, we compare these two simulations against a baseline where

Figure 14: Granular Banking Cycles



Notes: Impulse response functions to one-time negative shocks at $t = 5$ to the first (μ_ϵ) and third (λ_ϵ) moments of the distribution of idiosyncratic shocks, ξ. Red bars represent the case where only the banks belonging to the highest, 11th, permanent return type, κ, are affected. Blue bars represent the case where all banks in the distribution are affected.

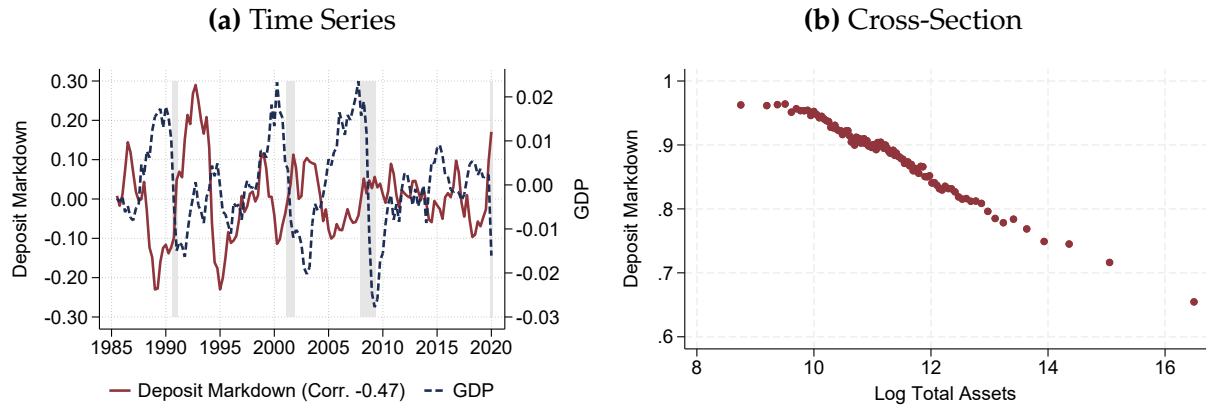
no shocks occur at $\mathbf{T}^*$, allowing us to derive the impulse response functions, as shown in Figure 14.

Results reveal that granular shocks (red bars) account for a significant portion—approximately 60%—of the total variation in real and financial variables induced by idiosyncratic shocks to all banks (blue bars). Given that granular shocks affect less than 10% of the banks in the economy, but explain more than half of total fluctuations caused by idiosyncratic shocks, the findings align closely with the canonical 80-20 Pareto principle (Gabaix, 2009). We conclude that idiosyncratic shocks to large banks alone can generate endogenous real and financial fluctuations, even in the absence of aggregate disturbances such as traditional TFP shocks. This result complements the findings of Carvalho and Grassi (2019) for non-financial firms and further supports the broader agenda of the granular hypothesis (Gabaix, 2011).

6.3 Deposit Market Power

Our baseline model assumes perfect competition among banks. However, a key feature of real-world data is the existence of bank market power, particularly in the deposit market (Drechsler et al., 2017, 2021; Egan et al., 2017; Wang et al., 2022). In this section, we depart

Figure 15: Deposit Market Power in the Data



Notes: Panel (a) plots the logged and HP-filtered estimated deposit markdown series. Panel (b) shows the binned scatter plot of the cross-sectional relationship between (log) total assets and estimated markdowns. All variables have been residualized from the time fixed effect. Source: U.S. Call Reports.

from the assumption of perfect competition and examine both empirically and within the model the quantitative implications of bank-level deposit market power. This allows us to explore how market power affects banking dynamics and the broader economy, offering a more realistic perspective on the role of banks in the financial system.

Deposit Markdowns in the Data We begin by presenting evidence of market power on the liability side of banks' balance sheets. Our empirical analysis utilizes Call Reports data. The methodology we follow is similar to that of De Loecker et al. (2020) and Corbae and D'Erasmo (2021). Specifically, we compute quarterly measures of bank-level deposit markdowns, denoted by $\mu^b_{j,t}$. The markdown $\mu^b_{j,t}$ is defined as the ratio of the marginal cost of raising a unit of deposits to a proxy for the interest rate on risk-free asset holdings. This methodology extends the measurement of credit markups from Corbae and D'Erasmo (2021) to the context of deposit markdowns. Details on the data are provided in Appendix A.1, while Appendix A.3 explains the estimation procedure.

Figure 15 presents key findings regarding the cyclical behavior of deposit markdowns and their relationship with bank size. In Panel (a), we display the cyclical component of deposit markdowns.[17] To construct this series, we compute the quarterly unweighted average of $\mu^b_t = \sum_j s_t \mu^b_{j,t}$, where s_t is the inverse of the number of banks in a given quarter.[18] The series is logged and filtered using the Hodrick-Prescott filter with the standard smoothness parameter of 1600, and the same treatment is applied to the U.S.

[17] Appendix A.4 also includes the raw, unfiltered series in levels. It also presents the components that constitute the markdown, i.e., the marginal cost and revenue from safe asset holdings.

[18] Appendix A.4 also confirms that the results hold when using quarterly weighted-averaging.

real GDP series. The results show that deposit markdowns are counter-cyclical, with a correlation of approximately -0.47 with filtered output, statistically significant at the 1% level.

Panel (b) of Figure 15 shows the cross-sectional relationship between markdowns and bank size, using binned scatterplots over 100 equally-sized bins of (log) total assets. The y-axes display bin-specific unweighted averages, and all variables have been residualized from the time fixed effect. These results indicate that market power on the liability side of banks' balance sheets is concentrated among larger banks, as markdowns decrease with bank size. In other words, larger banks charge lower deposit markdowns, highlighting a concentration of market power in the upper tail of the size distribution.

Before proceeding, we address a potential concern regarding the accuracy of our markdown measures. The "production function estimation" approach we follow might introduce a bias due to the assumptions about the prices and quantities of deposit goods. However, recent work by Grassi et al. (2022) on non-financial firms demonstrates that while this approach may yield an inaccurate *level* of aggregate markups, the behavior over time—particularly trends and *cyclical* components—remains highly correlated with true markup measures. Therefore, we can be confident that our cyclical components of markdowns are unbiased. Moreover, the Call Reports data provide precise balance sheet information on deposit returns and quantities, ensuring that a substantial portion of marginal costs is directly observable.

Markdowns in the Model Finally, we examine the quantitative implications of deposit market power within the context of our extended model. A detailed description of this model extension can be found in Appendix B.1. In this setup, households have two options for saving. First, they can save in mutual funds, denoted by M_t, which offer a guaranteed gross return R_t. Alternatively, households can save in bank deposits, which provide bank-specific non-contingent returns, $R_t^b(j)$. Importantly, bank deposits offer special liquidity services and contribute to the household's utility in an additively separable manner, consistent with the money-in-utility framework (Sidrauski, 1967; Walsh, 2010). The household's period utility function now takes the following form:

$$U(C_t, H_t, B_t) = \begin{cases} \frac{1}{1-\psi}\left(C_t - \chi_1 \frac{H_t^{1+\chi_2}}{1+\chi_2}\right)^{1-\psi} + \nu_b B_t & , \psi \neq 1 \\ \ln\left(C_t - \chi_1 \frac{H_t^{1+\chi_2}}{1+\chi_2}\right) + \nu_b B_t & , \psi = 1 \end{cases} \tag{28}$$

A new parameter, ν_b, governs the weight households place on deposit holdings in their

utility. Internalizing this liquidity effect, banks set the retail deposit rate, $R_t^b(j)$, by marking it down with $\mu_t^b(j) \leq 1$ over the risk-free rate R_t. Additionally, we assume that deposit franchises are imperfect substitutes, and the elasticity of substitution between them is denoted by $\theta_b > 0$. The first-order condition with respect to $b_t(j)$ leads to the following expression for deposit interest rates:

$$R_{t+1}^b(j) = R_{t+1}\left(1 - \left[\underbrace{\frac{U_{B,t}\left(C_t, H_t, B_t\right)}{U_{C,t}\left(C_t, H_t, B_t\right)}}_{\text{Marginal Liquidity Preferences}} \overbrace{\left(\frac{b_t(j)}{B_t}\right)^{\frac{1}{\theta_b}}}^{\text{Product Differentiation}}\right]\right) \tag{29}$$

where $U_{B,t}$ and $U_{C,t}$ denote marginal utility operators, and $R_{t+1} = \left[\beta \mathbb{E}_t \frac{U_{C,t+1}(C_{t+1}, H_{t+1}, B_{t+1})}{U_{C,t}(C_t, H_t, B_t)}\right]^{-1}$ 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 then be defined as follows:

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

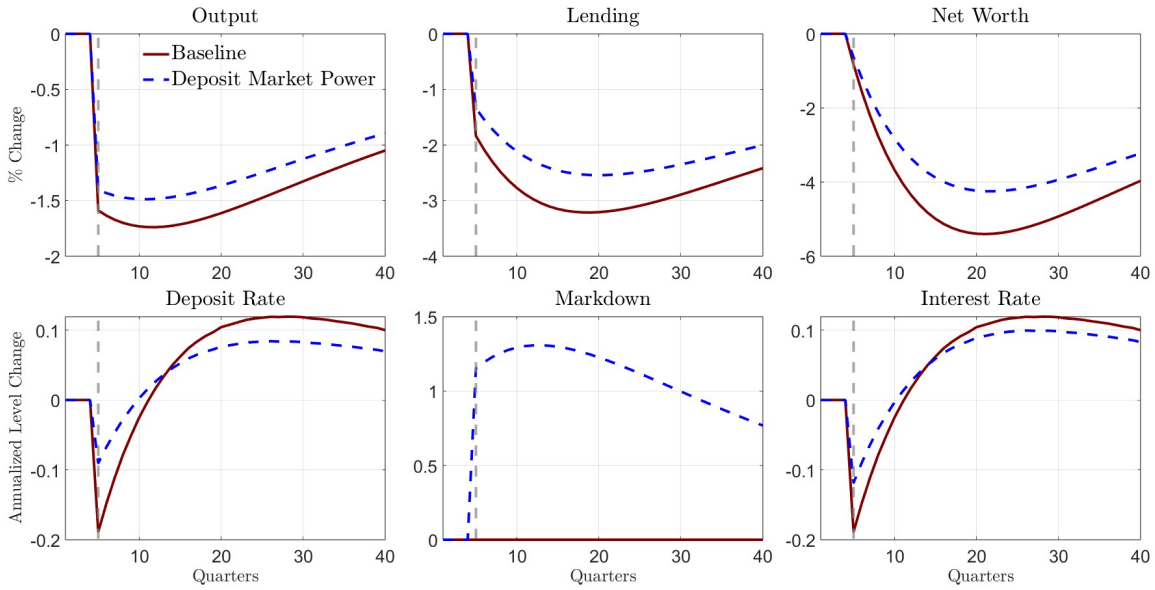
With the deposits-in-utility framework, marginal liquidity preferences vary with the aggregate state, causing markdowns to fluctuate over the business cycle. Moreover, since banks are heterogeneous and deposit franchises are not perfect substitutes, the markdowns banks choose depend explicitly on their idiosyncratic state vector. This leads to the emergence of a dynamic distribution of deposit market power in equilibrium.

We calibrate the two new parameters, ν_b and θ_b, as follows. In the data, the average deposit markdown—particularly in the latter years—is roughly 0.8 and we internally calibrate ν_b in order to target this moment. Parameter θ_b is internally calibrated to target the elasticity between bank-level markdowns and total assets, which in the data is -0.0067.[19] The bottom panel of Table 1 summarizes this parameterization approach.

Figure 16 illustrates impulse responses to a negative aggregate TFP shock under two scenarios: the baseline model and the extended model with deposit market power. The extended model successfully replicates the *counter-cyclical* behavior of deposit markdowns, in line with the data. The mechanism behind this result works as follows. In a bad aggregate state, bank lending and output both contract, but gradually, due to the slow response of net worth. Since the contraction is persistent, and follows a hump-shaped dynamic, consumption growth falls, leading to a fall in the risk-free real interest rate.

[19] We arrive at this value by running a panel regression with time and bank fixed effects of our estimated markdowns on bank assets based on the same Call Reports sample of banks as before.

Figure 16: Deposit Market Power in the Model



Notes: Impulse response functions to a one-standard deviation negative aggregate TFP shock in the baseline economy with ex-ante and ex-post bank heterogeneity and perfect banking competition, and in the extension with deposit market power.

Banks, however, exercise their market power in order to dampen deposit flight and lower the deposit rate *by less* than the risk-free rate, leading to a counter-cyclical markdown. In other words, the deposit interest rate is endogenously sticky and the spread between the risk-free rate and the deposit retail rate shrinks.

Overall, deposit market power *dampens* the effects of aggregate shocks relative to a representative-bank model with perfect competition. Banks, by raising markdowns in bad aggregate states, try to protect the demand for deposits, thereby allowing depositors (households) at the margin to smooth their response to shocks and preventing more severe deposit withdrawals and contractions in lending and real activity.

To summarize, in this section, we introduced two key features of U.S. banking data into our baseline model. First, we incorporated non-Gaussian idiosyncratic bank rate of return risk, characterized by pro-cyclical skewness. Second, we relaxed the assumption of perfect competition and introduced monopolistic competition in the deposit market. We demonstrated that both extensions significantly affect business-cycle fluctuations, highlighting the importance of these features in capturing real-world banking dynamics.

7 Additional Results and Robustness Tests

In the Appendix, we conduct several additional exercises, and here we offer a brief summary of select insights.

First, Appendix B.2 decomposes baseline impulse response functions into partial and general equilibrium effects. We find that the bulk of the transmission mechanism operates through the interest rate channel (direct effect), while the return on capital channel (indirect effect) plays a smaller, dampening role.

Second, in Appendix B.3, we relax the assumption of Pareto-distributed permanent types κ and assume instead that they follow a Normal distribution. In this specification, the mean and median values of κ across the eleven types are set to unity, as they are in the representative-bank benchmark. Despite the less dispersed and concentrated distribution of bank size, our amplification results remain robust due to the marginal propensity to lend (MPL) heterogeneity channel. Even in this setting, the average MPL exceeds that of the representative bank.

Finally, Appendix C provides computational details. Sections C.1 and C.2 describe our numerical algorithm and outline the accuracy checks performed to validate the results. Lastly, Section C.3 performs the Andrews et al. (2017) parameter sensitivity test.

8 Conclusions

We have developed a new tractable, dynamic stochastic general equilibrium framework with uninsured idiosyncratic bank rate of return risk, incomplete markets, and aggregate uncertainty. 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. Our approach breaks scale invariance, a feature that typically characterizes standard models with a representative intermediary. In this vein, the bank net worth distribution becomes a key dynamic object that has first-order effects on aggregate fluctuations. We introduce and characterize the marginal propensity to lend (MPL) as a sufficient statistic for the responses to aggregate shocks.

A version of our model with only incomplete markets and idiosyncratic shocks (ex-post heterogeneity) is isomorphic to a representative-bank benchmark. This result is due to a quasi-symmetry property of the model: the average MPL across the distribution is approximately identical to the MPL of a representative bank. Put differently, the equilibrium bank net worth distribution features a negligible degree of concentration relative to the data. Therefore, net worth accumulates among banks with a very similar

slope of the lending policy function, leading in turn to quasi-symmetric choices.

However, a model with *ex-ante* heterogeneity in bank profitability significantly breaks the isomorphism with the representative-bank economy. In that case, the net worth distribution features a much higher degree of concentration (at the top), closer to the data. This version of the model delivers significant real and financial amplification in response to aggregate shocks relative to a model with a representative bank. This is mainly due to the heterogeneity in the MPL statistic, with the average MPL in the model exceeding the one in the corresponding representative-bank economy. We then study two extensions of the baseline model that are rooted in micro banking data: counter-cyclical bank income risk and deposit market power. The former feature leads to further amplification of aggregate shocks, whereas the latter generates dampening and counter-cyclical deposit markdowns, a feature of the data replicated by our model.

Our 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.[20] 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.

[20] In Bellifemine et al. (2023) we study the monetary policy transmission mechanism in a Bewley Banks environment with nominal rigidities.

References

Abadi, Joseph, Markus Brunnermeier, and Yann Koby, "The Reversal Interest Rate," *American Economic Review*, August 2023, *113* (8), 2084–2120.

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

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

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," *American Economic Review*, 2024, *114.*

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 06 2017, *132* (4), 1553–1592.

Auclert, A., M. Rognlie, and L. Straub, "The Intertemporal Keynesian Cross," *Journal of Political Economy*, 2024.

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.*

Bellifemine, M., R. Jamilov, and T. Monacelli, "HBANK: Monetary Policy with Heterogeneous Banks," *CEPR Working Paper*, 2023, *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* (2), 435–439.

_ **and M. Gertler**, "Inside the Black Box: The Credit Channel of Monetary Policy Transmission," *American Economic Review*, 1995, *86* (4), 901–921.

_ **,** _ **, 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," *Working Paper*, 2023.

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).*

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

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.

Busch, 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).*

Calvo, G., A. Izquierdo, and E. Talvi, "Sudden Stops and Phoenix Miracles in Emerging Markets," *American Economic Review*, 2006, *96(2).*

Carlstrom, C. and T. Fuerst, "Agency Costs, Net Worth, and Business Fluctuations: A Computable General Equilibrium Analysis," *American Economic Review*, 1997, *87(5)*, 873–910.

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

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

Clerc, L., A. Derviz, C. Mendicino, S. Moyen, K. Nikolov, L. Stracca, J. Suarze, and A. Vardoulakis, "Capital regulation in a macroeconomic model with three layers of default," *International Journal of Central Banking*, 2015.

Coimbra, N. and H. Rey, "Financial Cycles with Heterogeneous Intermediaries," *The Review of Economic Studies*, 2023, *91(2).*

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** _ , "Banking Industry Dynamics Across Time and Space," *Working Paper*, 2023.

Corbae, Dean and Pablo D'Erasmo, "Capital Buffers in a Quantitative Model of Banking Industry Dynamics," *Econometrica*, 2021, *89(6).*

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.*

Dempsey, K., "Capital Requirements with Non-Bank Finance," *Review of Economic Studies*, 2024, *Forthcoming.*

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

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*.

Faccini, R., S. Lee, R. Luetticke, M. Ravn, and T. Renkin, "Financial Frictions: Macro vs Micro Volatility," *CEPR DP*, 2024, *15133*.

Farhi, Emmanuel and Jean Tirole, "Collective Moral Hazard, Maturity Mismatch, and Systemic Bailouts," *American Economic Review*, February 2012, *102* (1), 60–93.

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," *NBER Working Paper 27994*, 2021.

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

_ **, 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)*.

_ , _ **, and A. Queralto**, "Financial Crises, Bank Risk Exposure and Government Financial Policy," *Journal of Monetary Economics*, 2012, *59*, S17–S34.

Gilchrist, Simon and Charles P. Himmelberg, "Evidence on the role of cash flow for investment," *Journal of Monetary Economics*, 1995, *36* (3), 541–572.

Goldstein, Itay, Alexandr Kopytov, Lin Shen, and Haotian Xiang, "Bank Heterogeneity and Financial Stability," *Working Paper*, 2023.

Grassi, B., M. De Ridder, and G. Morzenti, "The Hitchhiker's Guide to Markup Estimatio," *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. and A. Krishnamurthy, "Intermediary Asset Pricing," *American Economic Review*, 2013, *103(2).*

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).*

Holmstrom, 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," *Journal of Economic Dynamics and Control*, 1993, *17.*

Imrohoroğlu, Ayse, "Cost of Business Cycles with Indivisibilities and Liquidity Constraints," *Journal of Political Economy*, 1989, *97* (6).

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

Janicki, H. and E. Prescott, "Changes in the Size Distribution of U. S. Banks: 1960 - 2005," *Federal Reserve Bank of Richmond Economic Quarterly*, 2004, *92* (4).

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

Kaplan, Greg and Giovanni L. Violante, "The Marginal Propensity to Consume in Heterogeneous Agent Models," *Annual Review of Economics*, 2022, *14* (1), 747–775.

_ , 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.

Krueger, D., K. Mitman, and F. Perri, "Chapter 11 - Macroeconomics and Household Heterogeneity," *Handbook of Macroeconomics*, 2016, *2*, 843–921.

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.*

Loecker, J. De, J. Eeckhout, and G. Unger, "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, 2020, *135(2).*

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.

Mendicino, C., K. Nikolov, J. Rubio-Ramirez, J. Suarez, and D. Supera, "Twin Default Crises," *Journal of Finance*, 2024, *Forthcoming.*

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

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

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*, 2023, *138.*

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

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.*

Stiglitz, J. and A. Weiss, "Credit Rationing in Markets with Imperfect Information," *American Economic Review*, 1981, *71(3).*

Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron, "Cyclical Dynamics in Idiosyncratic Labor Market Risk," *Journal of Political Economy*, 2004, *112* (3), 695–717.

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

Wang, O., "Banks, Low Interest Rates, and Monetary Policy Transmission," *Journal of Finance*, 2024, *Forthcoming.*

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).*

Werning, I., "Incomplete Markets and Aggregate Demand," *NBER Working Paper*, 2015, *21448.*

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

January 2025

Contents

A Empirical Appendix **2**
A.1 Data Details . . . 2
A.2 Data for Model Calibration . . . 4
A.3 Deposit Markdowns Estimation . . . 4
A.4 Additional Empirical Results . . . 6

B Model Appendix **10**
B.1 Model with Deposit Market Power . . . 10
B.2 Partial vs General Equilibrium Decomposition . . . 12
B.3 Normally Distributed Permanent Bank Heterogeneity . . . 14
B.4 Additional Model Results . . . 16

C Computational Details **18**
C.1 Numerical Algorithm . . . 18
C.2 Accuracy Checks . . . 22
C.3 Parameter Sensitivity Test . . . 24

A Empirical Appendix

A.1 Data Details

This section provides details on our data work. Table A1 summarizes all the series that are used throughout the paper. The main source of data 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 quarterly 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, loans, or total assets. 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.

Figure 1 plots the log of rank and the log of size for bank assets and deposits for the 4,000 largest (based on total assets) U.S. commercial banks for the first quarter of 2020. Our measure of total bank assets is the variable RCFD2170. Our measure of U.S. GDP growth, which is shown on Figure 15, is Gross Domestic Product obtained from the St. Louis Federal Reserve (FRED database). The series has been deflated with the CPI index, logged, and filtered with the Hodrick-Prescott filter (Ravn and Uhlig, 2002) under the usual smoothing parameter 1,600. Panel (a) shows the quarterly time series of markdowns that has been computed by taking equal-weighted averages, which is then logged and HP-filtered. The panel also reports the correlation coefficient of markdowns with respect to GDP growth; the value (negative 0.47) is statistically significant at the 1% level. Expansions and recessions are defined by the NBER criterion. Panel (b) shows a binned scatter plot with 100 equally-sized bins, with (log) total assets on the x-axis and markdowns in level on the y-axis. Dependent and independent variables have been residualized from the time fixed effect.

The Return on Loan (RoL) variable, which is displayed in Figure 12, is constructed as the ratio of interest income on loans (RIAD4010) to total loans (RCFD1400). We replace any missing values of total loans with loans net of unearned income and loss allowance. RoL growth is constructed by log-differencing at the bank level. In Panel (b), time series of the mean and skewness of RoL growth are computed with 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). The correlation

coefficients of the resulting objects with real GDP (which has been logged and HP-filtered) are 0.56 and 0.58, respectively, and both statistically significant at the 1% confidence level.

In Figure 6 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) with a similar interpretation.

Table A1: Variable Details and Sources

Variable	Details	Source
GDP	U.S. real Gross Domestic Product, chained 2012 dollars	FRED (GDPC1)
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 (RCFD1400)
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)
Deposit markdowns	Estimation procedure is detailed in Appendix A.3	Authors' calculation
Interest expense	Bank interest expenses on domestic deposits	Call Reports (RIAD4170-RIAD4172)
Non-interest expense	Bank non-interest expenses	Call Reports (RIAD4073+RIAD4093-RIAD4170-RIAD4180-RIAD4217)
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.

A.2 Data for Model Calibration

We now provide further details on the data that have been used for model calibration. The average 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 (Faccini et al., 2024). The average non-interest cost to loans ratio is computed as an unweighted average of the bank-level ratios. 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. Commercial bank assets and deposits Gini coefficients are computed for 2020:q1. The markdown elasticity of bank size has been estimated by running a panel regression of (log) markdowns on (log) total assets, a time fixed effect, and a bank fixed effect.

The standard deviation of output growth (σ_Y) has been computed from the logged and HP-filtered U.S. real GDP series. All correlation coefficients have been computed on pairs of variables that have been logged and HP-filtered. The correlation coefficient $\rho_{L,Y}$ is computed for the pair of output and real bank loans. The latter is our standard variable from the Call Reports. The coefficient $\rho_{N,Y}$ measures the correlation between output and real bank equity. $\rho_{LEV,Y}$ is the correlation coefficient between output and book leverage, defined as total loans divided by total equity. Finally, all panel variables have been winsorized at the 1% and 99% levels before the calculation of averages.

A.3 Deposit Markdowns Estimation

This section describes how we estimate deposit markdowns from U.S. bank-level data. We define the markdown of bank j in quarter t as:

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

where $c_{j,t}$ represents the marginal cost that bank j must incur in order to raise an extra unit of deposits and maintain the franchise, and $p_{j,t}$ is a proxy for the "safe revenue" that bank j collects in period t. 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. The variable $c_{j,t}$ is 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 common in the markup estimation literature:

$$\begin{aligned}\log(NIE_{j,t}) = {} & \alpha_i + \alpha_t + \beta_{l,1}\log(l_{j,t}) + \beta_{w,1}\log(w_{j,t}) + \beta_{q,1}\log(q_{j,t}) + \beta_{d,1}\log(d_{j,t}) \\ & + \beta_{l,2}\log(l_{j,t})^2 + \beta_{w,2}\log(w_{j,t})^2 + \beta_{q,2}\log(q_{j,t})^2 + \beta_{d,2}\log(d_{j,t})^2 \\ & + \beta_{l,w}\log(l_{j,t})\log(w_{j,t}) + \beta_{l,q}\log(l_{j,t})\log(q_{j,t}) + \beta_{w,q}\log(w_{j,t})\log(q_{j,t}) \\ & + \beta_{l,d}\log(l_{j,t})\log(d_{j,t}) + \beta_{w,d}\log(w_{j,t})\log(d_{j,t}) + \beta_{q,d}\log(q_{j,t})\log(d_{j,t}) \\ & + \beta_{t,1}t + \beta_{t,2}t^2 + \varepsilon_{j,t}\end{aligned} \tag{A1}$$

where $NIE_{j,t}$ is non-interest expenses, α_i and α_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, $q_{j,t}$ is total holdings of securities, $d_{j,t}$ denotes total domestic deposits, and t and t^2 denote respectively a linear and a quadratic time trend.[1] Further details on variables used are provided in Table A1

From (A1), it is straightforward to obtain marginal non-interest expenses as the derivative of non-interest expenses with respect to deposits:

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

The estimation of marginal non-interest income relies on the same procedure, with a caveat that we drop inputs from the right-hand side of the regression and the dependent variable is now (log) non-interest income:

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

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

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

Finally, marginal *net* non-interest expenses, *MNNIE*, are computed as the difference

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

between marginal non-interest expenses and marginal non-interest income:[2]

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

Panel (a) of Figure A1 plots the estimated markdown series, μ_t^b, computed as an unweighted quarterly average. Note that μ_t^b is now reported in levels and is unfiltered, unlike in Figure 15 where it was HP-filtered. Two observations are apparent from Figure A1. First, deposit markdowns do not exhibit any clear time-series trends. Instead, μ_t^b appears stationary and centered around 0.7-0.9, particularly in the latter years. The average markdown is also visibly counter-cyclical, in terms of its unconditional behavior. The markdown is most of the time below unity, implying the presence of deposit market power. In the 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).

Methodologically, our markdown estimation procedure complements the existing literature (Drechsler et al., 2017; Wang et al., 2022). The literature, by and large, measures deposit market power with deposit *spreads*, i.e., differences between retail deposit rates and the federal funds rate. Our approach builds on the influential study by De Loecker et al. (2020) and differs in at least two critical ways. First, our method is more sophisticated and captures more information than a reduced-form interest rate spread. In particular, it explicitly accounts for and computes the *marginal* cost of banks' deposits following the I-O literature. Second, our approach follows closely the production-function estimation of credit markups and asset market power (Corbae and D'Erasmo, 2021). As such, the methodological similarity allows for a direct comparison of the estimates and more robust takeaways. For example, we can conclude that while credit markups have been rising over time there has not been any noticeable trend in deposit markdowns.

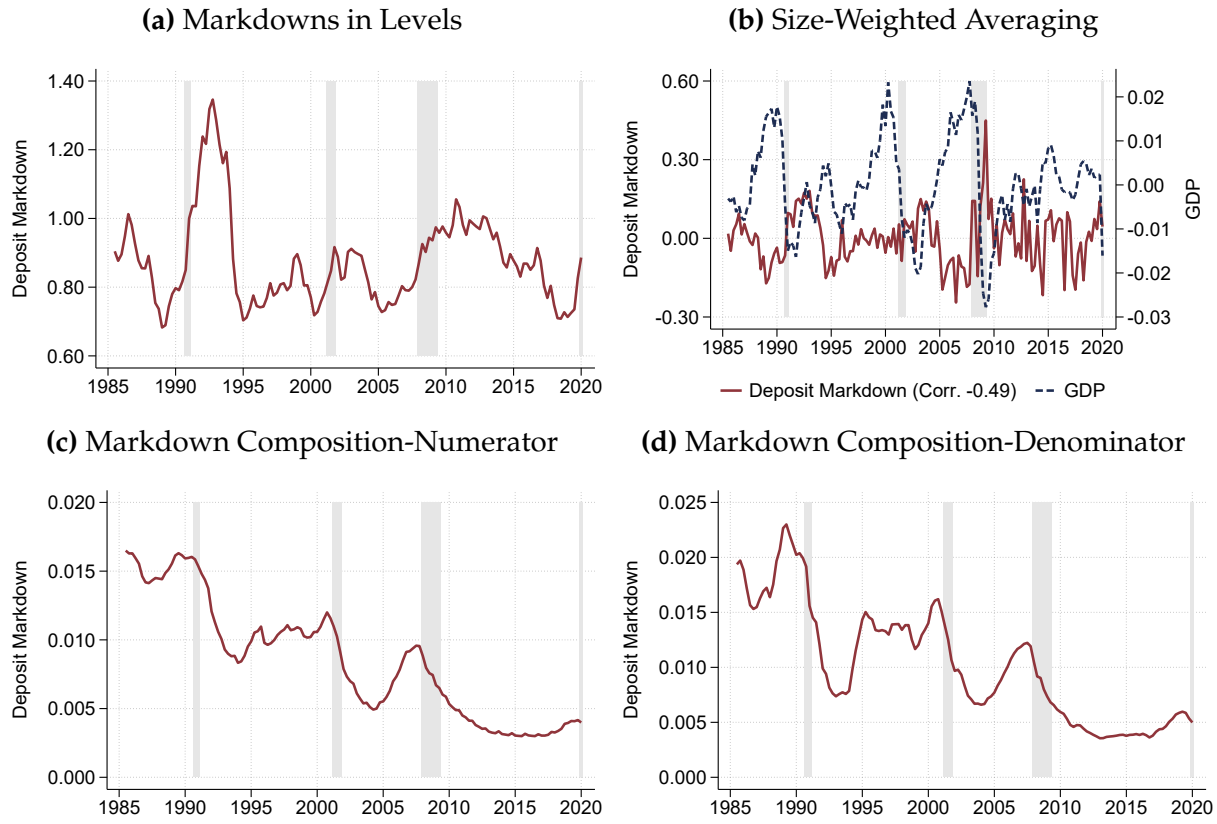
A.4 Additional Empirical Results

This section presents additional empirical results that supplement our findings in the main text.

Deposit Market Power Recall that, in the main text, our reported quarterly series of markdowns is the unweighted average of the panel. It is known in the literature that the aggregate properties of markups could be affected by how aggregation is performed. We

[2]To account for any influence of outliers, the resulting markdown series has been trimmed at the 2.5% and 97.5% levels. Finally, in our estimation sample we pre-remove all observations with a leverage ratio—defined as total assets over total equity—greater than 100, which constitutes less than 0.1% of the sample.

Figure A1: Deposit Markdown Robustness

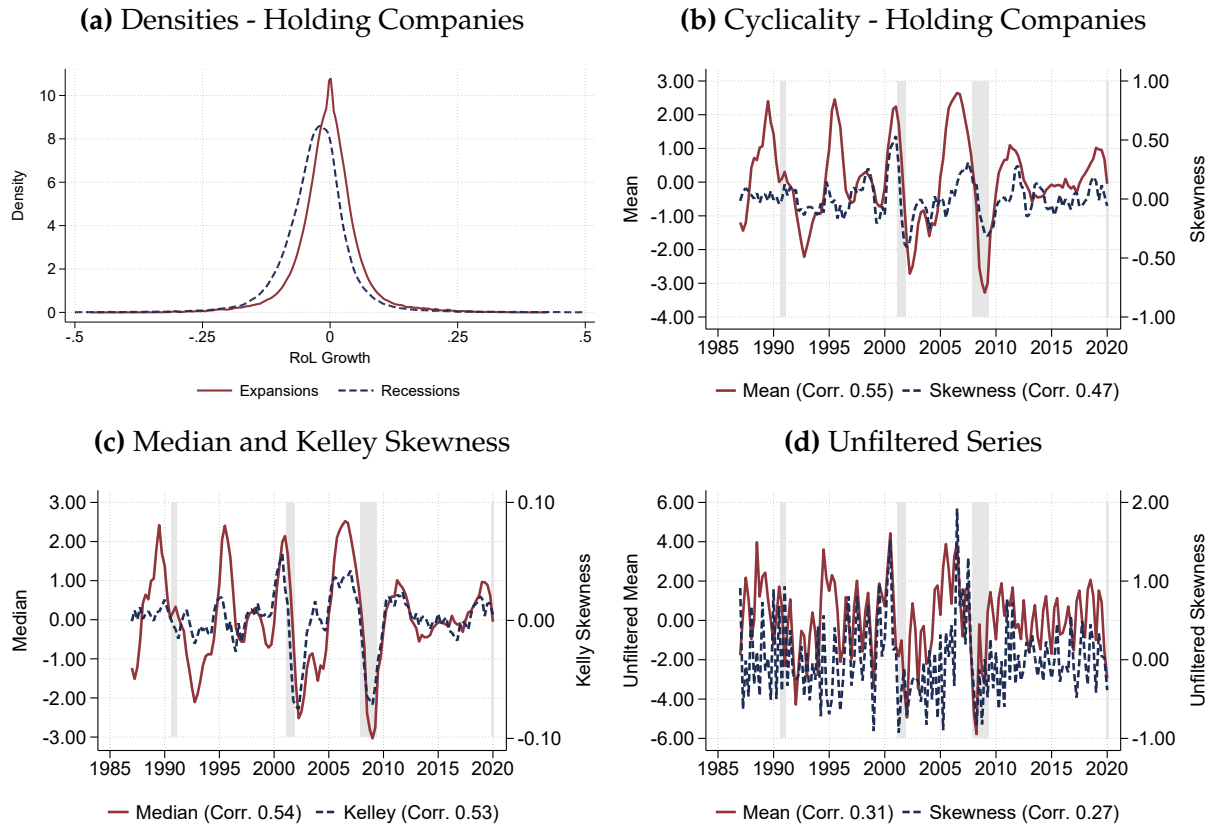


Notes: Panel (a) plots the time series of estimated deposit markdowns, computed by quarterly unweighted averaging and unfiltered. Panel (b) plots the time series of deposit markdowns, which has been computed by quarterly weighted averaging, logged, and HP-filtered. Bank-level total assets are used as weights. Panels (c) and (d), respectively, plot the numerator and denominator of the estimated markdown, in levels and computed as quarterly unweighted averages.

now compute a size-weighted average quarterly series and use total assets as a proxy for bank size. As usual, we HP-filter (logged) μ_t^b and report it alongside U.S. GDP in Panel (b) of Figure A1. Notably, while the level of μ_t^b changes slightly, its cyclical behavior is almost unchanged: it is still counter-cyclical with a pairwise correlation coefficient with respect to output equal to -0.49 and statistically significant at the 1% level.

For completeness, we also report the time series of the two key components that constitute the estimated markdown, i.e., the average marginal cost, c_t, and the average safe revenue, p_t. Panels (c) and (d) of Figure A1 present these objects as the numerator and denominator of the markdown, respectively. The ratio of the two produces the markdown in levels—shown on Panel (a)—exactly. We observe that both time series have been trending steadily over time. The decline of the marginal cost can be potentially explained by either the rise of efficiency in financial intermediation or by a selection effect (inefficient banks exit the market over time). Meanwhile, the decline in safe revenue rates

Figure A2: Bank Rate of Return Risk Robustness



Notes: Panels (a) and (b) plot densities of returns on loans, $\Delta r_{j,t}$, over the business cycle and the time series of the first and third moments of $\Delta r_{j,t}$ at the level of bank holding companies, respectively. Panel (c) plots the time series of the median and Kelley skewness for the baseline sample at the level of individual banks. Panel (d) plots the same time series as in Figure 12 except that they have been HP-filtered but not MA(4) filtered. Both series in Panels (b) and (c) have been HP-filtered and smoothed with an MA(4) filter.

is due to the general stylized fact of falling interest rates. While the two objects have been trending downward, the *ratio* of the two, however, has remained roughly stable over time, thus producing the markdown series as in Panel (a).

Countercyclical Income Risk The second check of robustness involves our second major extension of the baseline model: counter-cylicality of loan income risk. We now provide several robustness checks for that analysis. First, we focus on the level of aggregation. In the main text, we document that the first and third moments of the distribution of bank-level quarterly return on loan (RoL) growth are heavily procyclical. The level of aggregation in that finding is an individual bank. We now verify that this result holds at the level of bank holding companies. Figure A2 reports the densities of holding-level RoL growth in recessions and expansions (in panel (a)) and time series of the unweighted mean and skewness (in panel (b)). The takeaway from this check 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 recessions. Correlation coefficients of mean and skewness of holding-level RoL growth with output are 0.55 and 0.47, respectively, both statistically significant at the 1%.

Second, a potential concern with our measures of the first and third moments of the distribution of RoL growth is that the mean and statistical skewness can be affected by outliers. We now compute and present also the median of the distribution as a proxy for the first moment. In addition, as an alternative for the third central moment, Kelley skewness is often proposed in the literature (Guvenen et al., 2014) and is defined using deciles:

$$K_t = \frac{Q_{90,t} + Q_{10,t} - 2Q_{50,t}}{Q_{90,t} - Q_{10,t}}$$

where $Q_{x,t}$ is the x'th percentile or returns growth at time t. A positive value of K_t suggests that the right tail of the distribution of returns is wider than the left tail, and vice versa. Panel (c) of Figure A2 plots the time series of the median and the Kelley skewness (both HP-filtered and smoothed with an MA(4) filter) and shows that our results remain essentially unchanged: the first and third moments of the distribution of bank returns are procyclical and banks face greater downside risk in recessions. The correlation coefficients of the median and the Kelley skewness with U.S. output are 0.54 and 0.53, respectively, and both statistically significant at the 1% level.

Finally, for completeness we also report the unfiltered time series of the first and third moments of the distribution of bank returns. Our baseline approach involves smoothing the time series with a MA(4) process. One concern is that this could impact our cyclicality result. Panel (d) of Figure A2 shows the unfiltered mean and statistical skewness of RoL growth. Note that the two series have been still HP-filtered in order to isolate the cyclical component. We see that the raw series are still procyclical, with the correlations with U.S. GDP of 0.31 and 0.27, respectively. Both correlations are, moreover, statistically significant at the 1% level. Thus, while smoothing the series improves readability and interpretability of the data, it does not impact our conclusions.

Overall, we conclude that the above three robustness checks have validated the counter-cyclicality of bank return risk as a robust feature of the data.

B Model Appendix

B.1 Model with Deposit Market Power

In this section we provide a detailed setup of the model with bank market power in the deposit market. We begin with preferences. As before, the household derives utility from consumption as well as disutility from labor supply. Now, 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) or more generally to the money-in-utility framework (Sidrauski, 1967; Galí, 2008; Walsh, 2010). Mutual funds are risk-less investments but provide no liquidity utility. Both vehicles pay guaranteed, state non-contingent rates of returns. The 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 following form:

$$U(C_t, H_t, B_t) = \begin{cases} \frac{1}{1-\psi}\left(C_t - \chi_1 \frac{H_t^{1+\chi_2}}{1+\chi_2}\right)^{1-\psi} + \nu_b B_t & , \psi \neq 1 \\ \ln\left(C_t - \chi_1 \frac{H_t^{1+\chi_2}}{1+\chi_2}\right) + \nu_b B_t & , \psi = 1 \end{cases}$$

where ν_b, governs the weight households place on deposit holdings. Moreover, deposit products are now imperfect substitutes across banking franchises, still indexed by j, and assembled into the aggregate stock of deposits by the following aggregator:

$$B_t = \left[\int_0^1 b_t(j)^{\frac{\theta_b+1}{\theta_b}} dj\right]^{\frac{\theta_b}{\theta_b+1}} \tag{B3}$$

with $\theta_b > 0$ being the elasticity of substitution across deposit franchises. The flow budget constraint is given by:

$$C_t + \int_0^1 b_t(j)dj + M_t \leq R_t M_{t-1} + \int_0^1 R_t^b(j) b_{t-1}(j) + H_t W_t + \Pi_t + T_t \tag{B4}$$

where M_t are mutual fund holdings and $R_t^b(j)$ is the interest rate on deposits, which is now bank-specific. The first-order condition with respect to $b_t(j)$ yields a Lerner-type

formula for deposit interest rates:

$$R^b_{t+1}(j) = R_{t+1}\left(1 - \left[\underbrace{\frac{U_{B,t}\left(C_t, H_t, B_t\right)}{U_{C,t}\left(C_t, H_t, B_t\right)}}_{\text{Marginal Liquidity Preferences}} \overbrace{\left(\frac{b_t(j)}{B_t}\right)^{\frac{1}{\theta_b}}}^{\text{Product Differentiation}}\right]\right) \tag{B5}$$

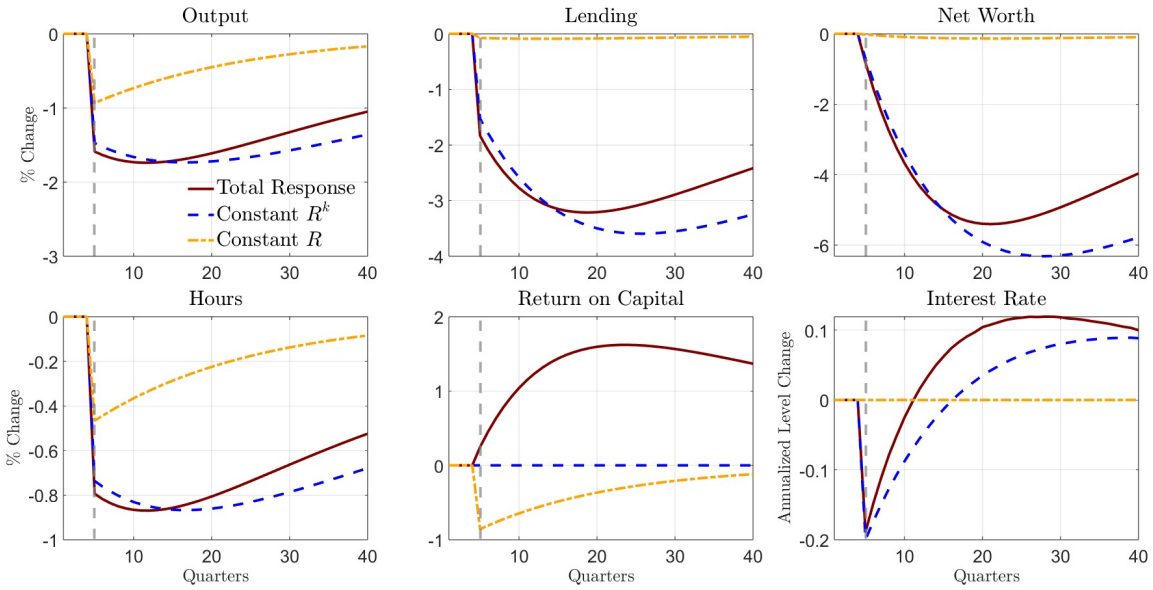
where $U_{B,t}$ and $U_{C,t}$ denote marginal utility operators, and $R_{t+1} = \left[\beta \mathbb{E}_t \frac{U_{C,t+1}(C_{t+1}, H_{t+1}, B_{t+1})}{U_{C,t}(C_t, H_t, B_t)}\right]^{-1}$ is the risk-free interest rate, which is pinned down by a first-order condition with respect to risk-less mutual fund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

$$\mu^b_t(j) = \frac{R^b_{t+1}(j)}{R_{t+1}} \leq 1$$

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 (B5) that the model can nest (a) perfect deposit market competition (if ν_b=0) or (b) a homogeneous markdown (if $\theta_b \to \infty$ for some finite $\nu_b > 0$). In this extended economy, the markdown is time-varying, heterogeneous, and proportional to relative deposit size.

We now discuss the appropriate changes to the dynamic banking problem. Banks take equation (B5) as given. Because households are not risk-neutral, and as long as θ_b is finite, banks also need to condition on aggregate consumption C_t and deposits D_t. The risk-free rate is taken as given as before. Conditional on C_t, D_t, and R_t, the partial equilibrium solution to the banking problem can thus be obtained. Bank-level markdowns are pinned down together with the distribution of relative deposit sizes, which are backed out via the balance sheet constraint once the lending policy function $\mathcal{L}(n, \kappa, \xi)$ is computed. The computation of general equilibrium proceeds as before. Appendix C.1 discusses how we solve this model extension numerically.

Figure B3: Partial vs General Equilibrium Decomposition



Notes: Impulse response functions to a one-standard deviation negative shock to aggregate TFP. Red straight lines depict baseline total responses. Yellow dotted lines isolate the direct effect with a time-invariant and exogenous interest rate, R_t. Blue dashed lines isolate the indirect effect with a time-invariant return on capital, R_t^k.

B.2 Partial vs General Equilibrium Decomposition

To further understand the mechanism underlying our model, this section decomposes baseline impulse response functions into partial (direct effect) and general equilibrium (indirect effect) channels. First, in order to identify the direct effect, we solve for the recursive equilibrium under the restriction that the real interest rate, R_t, is time-invariant and exogenously pinned down by time preferences. Shutting down the interest rate channel breaks down the feedback between the household sector and the dynamic banking problem. Second, to isolate the indirect effect, we solve for the recursive equilibrium under the assumption that the return on aggregate capital, R_t^k, is time-invariant and equals the return on capital implied by the version of the model without aggregate uncertainty. Shutting down the return on capital channel breaks down the feedback between banks and non-financial firms, i.e., the common component of bank portfolio returns is independent of firms' production decisions. This decomposition is similar to comparable approaches performed in the literature (Auclert, 2019).

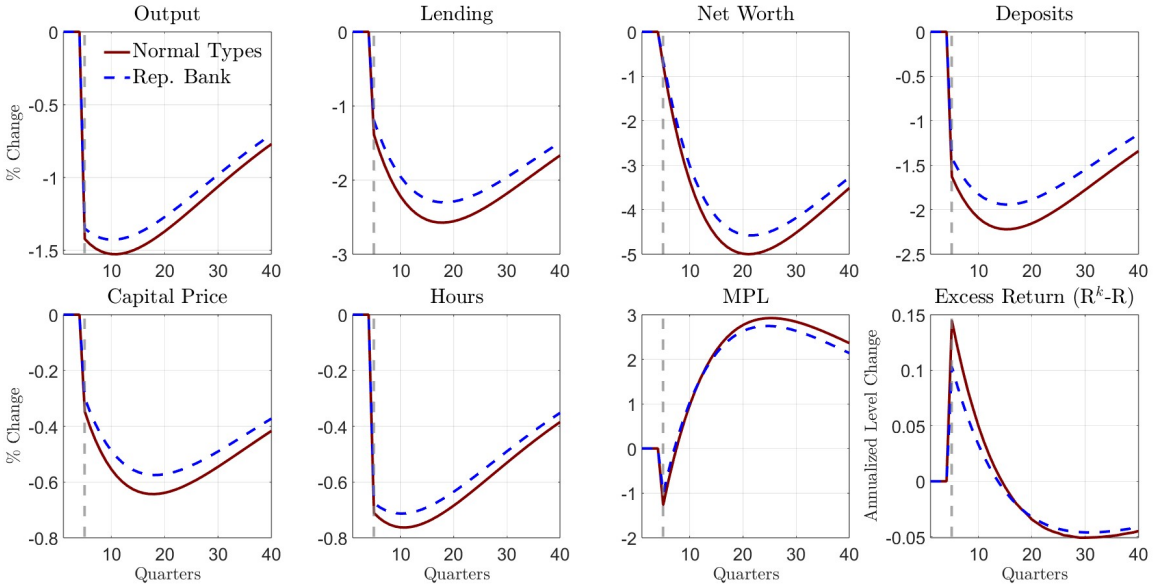
Figure B3 presents the results. We first plot baseline total responses to a one-standard deviation negative TFP shock. Relative to the baseline, shutting down the direct effect—as can be seen from the "Constant R" plots—more than halves the aggregate reaction from

output and hours and essentially nullifies the response of bank lending and net worth. A negative aggregate shock leads to a decline in output and consumption, which translates into a lower interest rate and higher excess returns. Normally, banks respond to tighter funding conditions by cutting lending supply to firms, which leads to a further decline in production and consumption, and so on. Quantitatively, it is apparent that the direct effect dominates the macroeconomic transmission mechanism in our model and plays a substantial amplifying role.

Second, the indirect effect ("Constant R^k" plots) channel represents a mild dampening force. Intuitively, the concavity of the production function raises the return on capital in recessions due to a decline in lending and capital. Absent this channel, banks face an even lower average portfolio return in bad aggregate states and, as a result, choose to cut lending by more.

The insights in this section are broadly consistent with prior literature findings in the context of non-financial firms (Ottonello and Winberry, 2020).

Figure B4: Normally Distributed Permanent Bank Heterogeneity



Notes: Impulse response functions to a one-standard deviation negative shock to aggregate TFP in the representative-bank economy and in the model with ex-ante and ex-post bank heterogeneity where permanent return types, κ, are Normally-distributed.

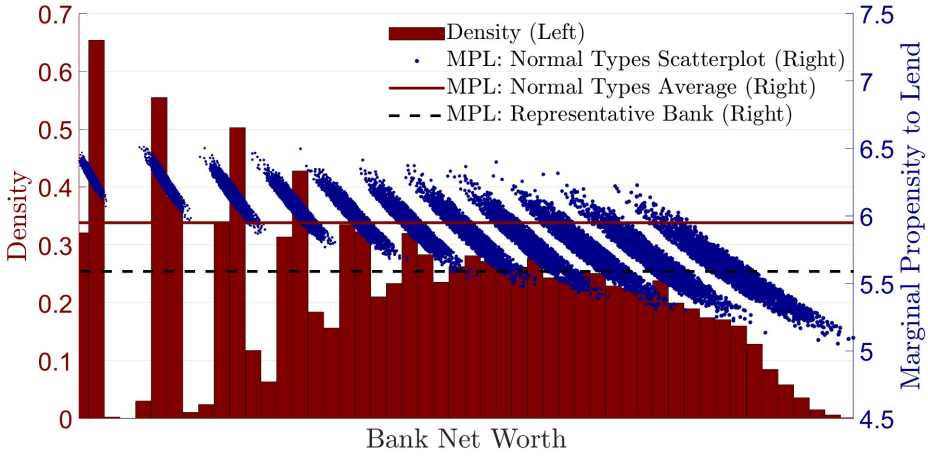
B.3 Normally Distributed Permanent Bank Heterogeneity

We have established that permanent bank rate of return heterogeneity can be a source of considerable amplification of aggregate shocks. In the baseline economy, we assumed that permanent bank types, $\kappa(j)$, follow a Power law of unit intensity, i.e., Zipf's law. The value of $\kappa(j)$ for the median, sixth, bank type has been normalized to unity. However, because of the right-skewed nature of the distribution of $\kappa(j)$, the average return is greater than unity. It is possible that our amplification result is obtained mechanically, considering that we have shown that the marginal propensity to lend (MPL)—the sufficient statistic for aggregate responsiveness to shocks—is increasing in κ and there is a small number of banks with abnormally high permanent profitability.

In this section, we address this issue by solving for the recursive equilibrium of the model with $\kappa(j)$ drawn from a Normal distribution with volatility set to the empirical cross-sectional standard deviation of bank returns' fixed effects[3]. As a result, both the median and the average permanent return are equal to unity, as is the case by design in the representative-bank special case.

[3]Specifically, when estimating a linear panel fixed-effect model with AR(1) disturbances on returns of loans (RoL) in order to parameterize the idiosyncratic income process, we extract the fixed effects and compute their standard deviation.

Figure B5: MPL Heterogeneity with Normally Distributed Bank Types

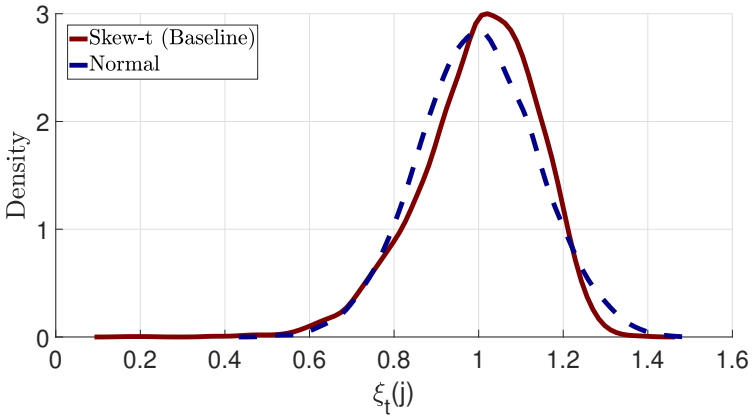


Notes: Marginal propensity to lend (MPL) across the distribution of bank-level net worth in the recursive equilibrium of the model with ex-ante and ex-post heterogeneity and where permanent return types, κ, are Normally-distributed. Straight and dashed horizontal lines refer to average MPLs in the economy with heterogeneity and the representative-bank benchmark, respectively. The left y-axis shows the density of the distribution and the right y-axis shows the levels of the MPL, in percentage points.

Figure B4 presents impulse response functions based on the model with normally distributed bank types. Across both financial and economic aggregates, we notice that the amplification result is maintained. Relative to the representative-bank economy, the model with bank heterogeneity still delivers starker recessions. In order to understand why, we revisit the distribution of marginal propensities to lend (MPL). Figure B5 shows that the MPL of the representative bank is still lower than the average MPL of the model with heterogeneous banks, even when permanent heterogeneity is drawn from a symmetric density.

One may wonder why normally-distributed types $\kappa(j)$ are not adopted as the baseline specification. The reason is that—as can be clearly seen from Figure B5—this version of the model fails to deliver a plausible, concentrated distribution of bank net worth. As a result, all the practically and policy relevant questions related to granular bank failures or the too-big-to-fail externality would not be possible to examine in such an environment. Since Pareto-distributed permanent heterogeneity does not skew our results and delivers a realistic bank size distribution, we maintain it as the baseline.

Figure B6: Hansen's Skew-t Density

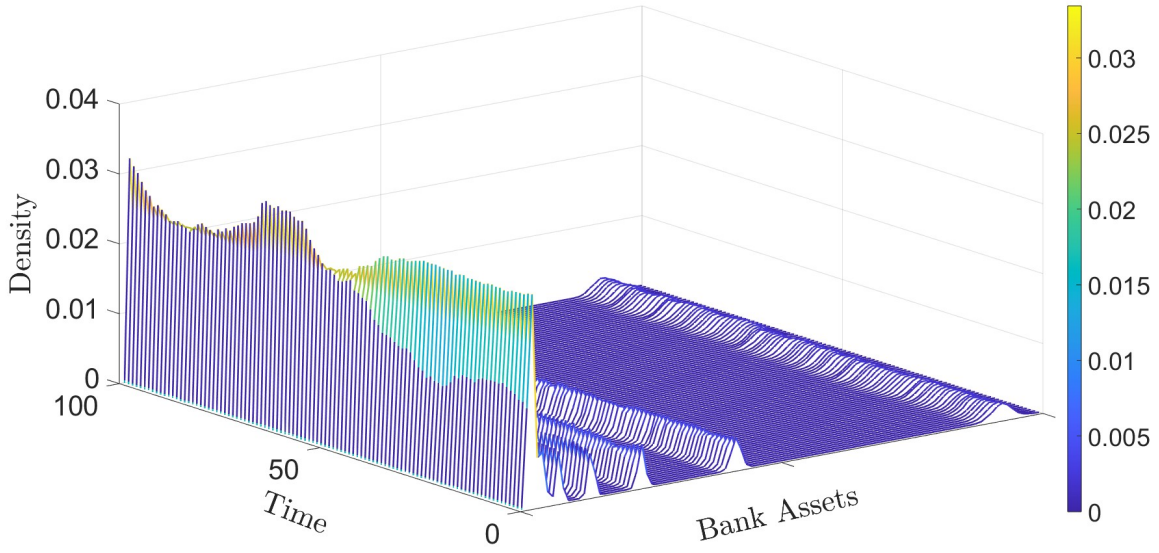


Notes: The Hansen (1994) Skewed-t density (solid) vs the Normal density (dashed).

B.4 Additional Model Results

In this section we present additional model results to supplement the main text. Figure B6 plots the density of idiosyncratic bank rate of return risk, $\xi_t(j)$, in two versions of the model. First, the baseline case with Normally-distributed shocks. Second, the extension of the model with countercyclical returns and where $\xi_t(j)$ are drawn from the Hansen (1994) Skew-t density. The major difference is the introduction of a sharp, prolonged left-tail. Left-skewness appears whenever $\lambda_{\epsilon,t} < 0$. For the special case of $\lambda_{\epsilon,t} = 0$, the Skew-t collapses exactly to the Normal distribution.

Figure B7: Dynamics of the Banking Distribution



Notes: A waterfall graph depicting the dynamics of the distribution of bank assets over time.

While in the main text we have presented the distribution of bank size in a given period of time, we now showcase how the distribution evolves over time in our baseline model with both ex-ante and ex-post bank heterogeneity. To this end, we present a so-called *waterfall* graph that plots the fully dynamic density of bank assets. For readability, we only show 100 periods from our full simulation. Figure B7 presents the plot. It is clear from the Figure that (i) the distribution is always highly right-skewed and features a small number of very large intermediaries in the right tail; (ii) the distribution is very dynamic and moves around over time in response to both aggregate and idiosyncratic disturbances while preserving the basic shape.

C Computational Details

In this Section we discuss our computational solution method, provide checks of numerical accuracy, and run a test of parameter sensitivity and model identification.

C.1 Numerical Algorithm

This section provides details on the numerical algorithm that we use to solve our model. The algorithm is similar to the canonical method of Krusell and Smith (1996, 1998). The model that is closest to ours in terms of the computational approach is the Krueger et al. (2016) framework that includes ex-ante and ex-post household heterogeneity, counter-cyclical income risk, and aggregate uncertainty.

Baseline Model We begin with the baseline economy that features ex-ante and ex-post bank heterogeneity, a-cyclical income risk, and perfect competition. Our model features an endogenous aggregate state variable—the distribution of bank net worth Γ_t—which we track with the first moment, denoted by $\bar{N}_t$. Conditional on $\bar{N}_t$, and using the lending policy function, one can obtain the average level of bank claims, $\bar{L}_t$, capital, K_t, and the price of capital, Q_t. These objects are sufficient to pin down the rate of return on capital, r_t^k, which is then taken as given by the banks' decision problem.

The goal is to solve for the equilibrium law of motion of the distribution, F^*, as well as to approximate the functions $V(n, \kappa, \xi; \bar{N}, A)$ and $\mathcal{L}(n, \kappa, \xi; \bar{N}, A)$. We solve the partial-equilibrium problem of banks with value function iteration on a grid of points in the $(n, \kappa, \xi, \bar{N}, A)$ plane and by using the modified Akima interpolation for points that are not on the grid.

In order to eventually solve the baseline model with aggregate uncertainty, we proceed incrementally with the following steps:

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

   In order to construct the right grid for $\bar{N}$, we first solve for a *stationary equilibrium* and determine the stationary value of aggregate bank net worth N^{ss}. In other words, we shut down aggregate uncertainty by normalizing the value of aggregate productivity, A, to unity. This version of the model resembles a framework similar to Aiyagari (1994) with endogenous capital accumulation and financial frictions.

   We set up an exponential grid for bank-level net worth n with 36 points over the interval $[\underline{n}, \overline{n}]$ such that $\underline{n}$ is a small positive number and $\overline{n}$ is a non-binding upper limit. The distribution of permanent returns, κ, is discretized by drawing a large

array of random numbers from a Pareto I density with the shape parameter $\alpha_\kappa = 1$. The constant k_m has been normalized such that the distribution's median value is unity. Then, we define 11 permanent return *types* with the respective percentiles of the drawn values such that the 6th type is unity. We discretize the distribution of idiosyncratic rate of return shocks, ξ, with the Tauchen (1986) method and 5 points centered around unity.

We solve for the stationary equilibrium in three steps. First, for a given guess of aggregate quantities and prices, the partial-equilibrium problem of banks is solved with value function iteration. Second, using the obtained candidate policy functions we run a simulation with 2,000 periods. Third, a new vector of aggregate quantities and prices is constructed by taking averages of the simulated time series. If the absolute value difference between the old and the new values for aggregate loans, $\bar{L}$, is less than a specified tolerance level of 10^{-3}, the algorithm stops. If not, it resumes with another iteration until convergence is obtained. Upon completion, the equilibrium value of aggregate net worth, N^{ss}, is stored.

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

- *Preliminaries.*

  Having solved for the stationary equilibrium, we build an equally-spaced grid for aggregate bank net worth, $\bar{N}$, with 4 points and centered around the stationary steady-state value N^{ss}. We assume that aggregate productivity takes on two values: $\{A_H, A_L\}$ with $A_H - A_L = \Delta_a$. The probability matrix that governs transitions across aggregate states is $\boldsymbol{\pi_a}$. Both objects are chosen as per the calibration Table (1).

- *Law of motion of the distribution.*

  A key aspect of our numerical solution is how we deal with the endogenous, time-varying distribution of banks, Γ_t. We follow Krusell and Smith (1998) and assume that banks build limited-information forecasts based on the first moment of the distribution of net worth, $\bar{N}$. The limited-information aggregate state vector is thus given by $(\bar{N}, A)$. We conjecture that the law of motion for aggregate net worth is log-linear, as first mentioned in the main text:

$$
\begin{aligned}
A = A_L: \quad & \log \bar{N}' = \beta^N_{L,0} + \beta^N_{L,1} \bar{N} \\
A = A_H: \quad & \log \bar{N}' = \beta^N_{H,0} + \beta^N_{H,1} \bar{N}
\end{aligned}
$$

The forecast for the first moment of the distribution of bank claims, $\bar{L}$, conditional on net worth, is as follows:

$$A = A_L : \quad \log \bar{L}' = \beta^L_{L,0} + \beta^L_{L,1} \bar{N}$$
$$A = A_H : \quad \log \bar{L}' = \beta^L_{H,0} + \beta^L_{H,1} \bar{N}$$

The fixed point for F^* is given by the vector $\boldsymbol{\beta} = \left\{\beta^{N*}_{L,0}, \beta^{N*}_{L,1}, \beta^{N*}_{H,0}, \beta^{N*}_{H,1}, \beta^{L*}_{L,0}, \beta^{L*}_{L,1}, \beta^{L*}_{H,0}, \beta^{L*}_{H,1}\right\}$ which is sufficient to characterize the dynamics of the distributions of net worth and claims and, as a result, every other aggregate variable of the economy. Also denote by $\boldsymbol{\beta}_N = \left\{\beta^{N*}_{L,0}, \beta^{N*}_{L,1}, \beta^{N*}_{H,0}, \beta^{N*}_{H,1}\right\}$ the subset of the law of motion that corresponds to net worth.

We now perform the following steps until convergence:

(I) *Solve the banking problem.*

Start with some initial guesses for the law of motion of the distribution, $\boldsymbol{\beta}^{(i)}$, and the interest rate, $R^{(i)}$, denoting by superscript (i) the iteration count. Conditional on the return on aggregate capital, $R^{K(i)}$, that is implied by $\boldsymbol{\beta}^{(i)}$, solve the banking problem and obtain the candidate value function, $V^{*(i)}$, and the corresponding policy function, $\mathcal{L}^{(i)}$.

(II) *Montecarlo simulation.*

Simulate the model with a panel of 2,000 banks for 2,000 periods. In each period, compute end-of-period average net worth, $\bar{N}^{(i)}_t$, and claims, $\bar{L}^{(i)}_t$, using the just-obtained policy functions. Using the sequence for loans, compute the sequences for aggregate capital, $K^{(i)}_t$, and the price of capital, $Q^{(i)}_t$, using the non-financial firms block as well as labor supply, $H^{(i)}_t$, from GHH preferences. Compute the path of output, $Y^{(i)}_t$, using the aggregate production function and consumption, $C^{(i)}_t$, from the goods market clearing condition. Using the consumption path, construct the stochastic discount factor, $\Lambda^{(i)}_t$, and the risk-free rate, $R^{(i)}_t$.

(III) *Update laws of motion.*

Using the simulated data, run linear OLS regressions of (log) $\bar{N}^{(i)}_t$ on (log) $\bar{N}^{(i)}_{t-1}$ and a constant, having discarded the first 500 periods, and conditional on the aggregate state of the economy. Similarly, run regressions of (log) $\bar{L}^{(i)}_t$ on (log) $\bar{N}^{(i)}_{t-1}$ and a constant. Obtain the new candidate for the forecasting rule $\boldsymbol{\beta}^{(i+1)}$. Compute average values of the interest rate in good and bad aggregate states, $R^{(i+1)}_{A_H}$ and $R^{(i+1)}_{A_L}$.

(IV) *Convergence test.* Compute the Euclidian norm between $\boldsymbol{\beta}_N^{(i)}$ and $\boldsymbol{\beta}_N^{(i+1)}$. If the difference is below a chosen level of tolerance, 10^{-2}, the algorithm concludes. Otherwise, slowly update the forecasting rule as follows: $\boldsymbol{\beta}^{(i+1)} = \omega\boldsymbol{\beta}^{(i+1)} + (1-\omega)\boldsymbol{\beta}^{(i)}$ with $\omega = 0.1$ and proceed with step (I) again until convergence is achieved.

Representative Bank Case The solution of the representative-bank economy is substantially less complex and time-consuming. We discard ex-ante and ex-post bank heterogeneity. The state vector becomes $(n; \bar{N}, A)$. Note that bank-level net worth, n, is still a state due to the scale variance property of the model. The approach otherwise consists of the same steps as before.

Countercyclical Returns Our extension with counter-cyclical rate of return risk features a non-Gaussian distribution of shocks and pro-cyclical skewness. To discretize that distribution, we use a variant of the Tauchen (1986) approach which has been modified to handle the case of non-Gaussian shocks. First, we use the standard method to obtain a matrix of transition probabilities for a stochastic process ξ with volatility σ_ξ and persistence ρ_ξ 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 dyad $\{\lambda_\epsilon, \eta\}$. 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 $\lambda_\epsilon = 0$.

In this case, transitory risk, ξ, becomes aggregate state-dependent. We assume that high-productivity states, A_H, are characterized by ξ that is drawn from a Skew-t distribution with $\lambda_\epsilon = 0$ while the low-productivity state, A_L, features ξ drawn from a Skew-t distribution with $\lambda_\epsilon = -0.5$. That is, banks face de-facto Gaussian idiosyncratic shocks in normal states and left-skewed non-Gaussian shocks with greater downside risk in bad states.

The above modification aside, the numerical algorithm is unchanged.

Deposit Market Power In the extension of the model with deposit market power, bank-level deposit rates, $R_{t+1}^b(j)$, depend on (i) aggregate consumption, C_t, and (ii) aggregate deposits, $\bar{B}_t$, which, as before, are tracked by the first moment of the distribution. We employ projection methods similar to the above and construct log-linear forecasts of the

Table C2: Equilibrium Forecasting Rules and Accuracy Metrics

		F^*		R^2		Mean Error	SD Error
		A_L	A_H	A_L	A_H	A_L	A_H
$\bar{N}$	β_0^N	0.184	0.200	0.999	0.999	0.617%	0.475%
	β_1^N	0.938	0.938				
$\bar{L}$	β_0^L	3.435	3.476	0.962	0.934	0.475%	0.361%
	β_1^L	0.496	0.494				

Notes: This table reports equilibrium laws of motion of average net worth, $\bar{N}$, and average claims, $\bar{L}$. It also shows the values of R^2 from linear OLS regressions that are run in Step (III) of the numerical algorithm. The last two columns report results from the Den Haan (2010) accuracy test and show means and standard deviations of percentage differences between the actual model-implied time series of $\bar{N}_t$ and $\bar{L}_t$ and the series that are constructed with the equilibrium forecasting rule, $\boldsymbol{\beta}$.

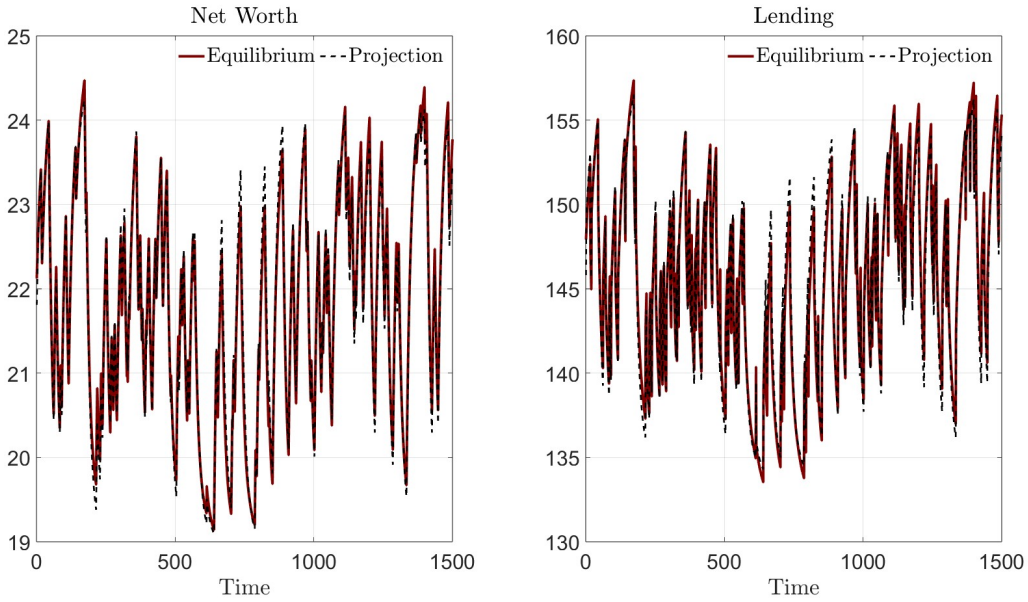
two objects conditional on aggregate net worth $\bar{N}_t$. Denote with $\boldsymbol{\beta}_B$ and $\boldsymbol{\beta}_C$ the vectors of coefficients that summarize the respective laws of motion. Step (III) of the numerical algorithm now includes running regressions of (log) $\bar{B}_t^{(i)}$ and (log) $\bar{C}_t^{(i)}$ on (log) $\bar{N}_t^{(i)}$ and a constant. The dynamic banking problem takes the new laws of motion as given and pins down the distribution of bank-level markdowns: the wedge between the risk-free rate R_t and $R_{t+1}^b(j)$. The numerical algorithm is otherwise unchanged.

Our numerical algorithm, 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 Γ_t, specifically by leveraging neural networks (Nuno et al., 2023). However, we do not pursue these extensions here because our solution is already accurate enough.

C.2 Accuracy Checks

In order to test whether our algorithm is accurate, we perform two checks. First, it is important to verify that the equilibrium projection rules, $\boldsymbol{\beta}$, can explain a high percentage of the variation of model-implied aggregates. Table C2 reports the R^2 from the regressions that we run in Step (III) of the numerical algorithm above. The law of motion of the banking distribution is approximated very accurately—in both good and bad aggregate states—as can be seen from the first row, with the R^2 of above 99.9%. This confirms that

Figure C8: Equilibrium and Forecasted Aggregates



Notes: This figure plots the actual model-implied paths of $\bar{N}_t$ and $\bar{L}_t$ and the series that are constructed with the equilibrium forecasting rule, $\boldsymbol{\beta}$.

approximating the distribution with the first moment does quite well in terms of capturing the dynamics of the full distribution of net worth. This result also complements the earlier observation that the law of motion of aggregate net worth is approximately linear, as has been shown in Figure 2. For the projection of aggregate bank loans, $\bar{L}_t$, on net worth, we obtain R^2 of around 0.93-0.96, somewhat lower but still very high values by literature standards.

A second accuracy test encourages us to go beyond simply reporting 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 $\boldsymbol{\beta}$. Figure C8 plots actual and forecasted values of aggregate net worth and loans. It is clear that projected values track the actual ones very closely. Table C2 also reports the mean and standard deviation of the percentage difference between actual and projected values for the two variables. Errors are very low, to the order of around just half of a percentage point with a standard deviation of around 0.4%, on average.

Table C3: Parameter Sensitivity Test

	Hours	Non-interest Ratio	Price	Leverage	Gini	MPL
χ_1	-0.30%	-0.21%	0.00%	-0.01%	0.00%	-0.01%
ζ_1	-0.09%	0.07%	-0.13%	0.12%	-0.08%	0.10%
a	0.00%	-0.21%	-0.99%	-0.01%	0.00%	-0.01%
λ	0.01%	-0.23%	0.02%	-1.34%	-0.01%	-1.32%
α_κ	-0.16%	-1.98%	-0.21%	-0.15%	-0.36%	-0.14%
σ_ξ	0.00%	-0.20%	0.00%	-0.05%	0.00%	-0.05%

Notes: This table reports results from the Andrews et al. (2017) parameter uncertainty test. Values represent local elasticities of model moments (shown in columns) with respect to marginal changes in parameters (shown in rows).

C.3 Parameter Sensitivity Test

To help understand the directions of identification in our model calibration, we perform a version of the Andrews et al. (2017) parameter sensitivity test on the stationary equilibrium. In particular, we want to gauge the responsiveness of key moments of the model to changes in select parameters. The six moments of interest are, in order: hours worked, the average non-interest expense-to-loans ratio, the price of capital, the average leverage ratio (defined as claims over net worth), the Gini coefficient for the distribution of bank claims, and the average marginal propensity to lend (MPL). The six parameters that we consider are, in order: χ_1 which gauges labor disutility, ζ_1 which controls non-interest expenses, parameter a of the production function of non-financial firms, the leverage constraint parameter λ, α_κ which determines the intensity of the Pareto I distribution from which permanent bank types κ are drawn, and σ_ξ, which is the volatility of the idiosyncratic rate of return process.

Results are shown in Figure C3 in the form of local elasticities of moments (shown in columns) with respect to parameters (shown in rows). Parameter χ_1, which controls preferences for leisure, has a large negative impact on the hours worked. Increasing ζ_1 mildly raises non-interest expenses. Raising a lowers the price of capital almost one-to-one, which is intuitive. The leverage parameter λ has a large negative impact on leverage and the MPL. Raising α_k reduces the skew in the distribution of permanent types and, as a result, lowers equilibrium size concentration as evidenced by the negative impact on the Gini coefficient. Interestingly, it appears that α_k is the most informative parameter for all moments, consistent with our broader conclusion that permanent bank heterogeneity has first-order impacts on the recursive equilibrium. Finally, the impact of σ_ξ is quantitatively very mild. Overall, we conclude that all the above patterns are logical and consistent with

the rest of the findings in the paper.

References

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

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 06 2017, *132* (4), 1553–1592.

Auclert, Adrien, "Monetary Policy and the Redistribution Channel," *American Economic Review*, June 2019, *109* (6), 2333–67.

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.*

Corbae, Dean and Pablo D'Erasmo, "Capital Buffers in a Quantitative Model of Banking Industry Dynamics," *Econometrica*, 2021, *89(6).*

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

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.*

Faccini, R., S. Lee, R. Luetticke, M. Ravn, and T. Renkin, "Financial Frictions: Macro vs Micro Volatility," *CEPR DP*, 2024, *15133.*

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

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

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

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

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

Krueger, D., K. Mitman, and F. Perri, "Chapter 11 - Macroeconomics and Household Heterogeneity," *Handbook of Macroeconomics*, 2016, *2*, 843–921.

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.

Loecker, J. De, J. Eeckhout, and G. Unger, "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, 2020, *135(2).*

Nuno, G., 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).*

Ravn, M. O. and H. Uhlig, "On Adjusting the Hodrick-Prescott Filter for the Frequency of Observations," *The Review of Economics and Statistics*, 2002, *84.*

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

Tauchen, George, "Finite state markov-chain approximations to univariate and vector autoregressions," *Economics Letters*, 1986, *20* (2).

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

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).*