DISCUSSION PAPER SERIES

DP17129

HBANK: Monetary Policy with Heterogeneous Banks

Marco Bellifemine, Rustam Jamilov and Tommaso Monacelli

INTERNATIONAL MACROECONOMICS AND FINANCE

MONETARY ECONOMICS AND FLUCTUATIONS

CEPR

ISSN 0265-8003

HBANK: Monetary Policy with Heterogeneous Banks

Marco Bellifemine, Rustam Jamilov and Tommaso Monacelli

Discussion Paper DP17129
Published 22 March 2022
Submitted 21 March 2022

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- International Macroeconomics and Finance
- Monetary Economics and Fluctuations

HBANK: Monetary Policy with Heterogeneous Banks

Abstract

We study how bank heterogeneity and market power shape the transmission of monetary policy. In the data, following a monetary contraction, large banks lower their credit mark-ups and quantities significantly more and raise their deposit mark-ups significantly less than small banks. We interpret our findings in a Heterogeneous Bank New Keynesian (HBANK) model, featuring permanent ("skill") and stochastic ("luck") bank returns heterogeneity, incomplete markets, variable asset and deposit market power, and nominal rigidities. In this setup, the aggregate effects of monetary policy shocks depend explicitly on the endogenous distribution of banks' net worth and the competitive structure of asset and deposit markets. Our main findings are three-fold. First, the model matches both the aggregate and the cross-sectional conditional responses to monetary policy shocks estimated from the data. Second, due to the marginal propensity to lend being increasing in profitability, permanent returns heterogeneity considerably strengthens the effectiveness of monetary policy shocks. Third, market power amplifies the effects of monetary policy shocks, whereas the opposite holds for deposit market power. This stems from credit mark-ups of large intermediaries being strongly conditionally pro-cyclical, while deposit mark-ups being counter-cyclical. When calibrated to U.S. bank-level data, the full model delivers substantial amplification of monetary shocks.

JEL Classification: E44, E51, E52

Keywords: monetary policy, Financial Intermediation, market power, Bank Heterogeneity

Marco Bellifemine - m.bellifemine@lse.ac.uk
London School of Economics

Rustam Jamilov - jamilovrustam@gmail.com
All Souls College, University of Oxford

Tommaso Monacelli - tommaso.monacelli@unibocconi.it
Bocconi University and CEPR

HBANK:
Monetary Policy with Heterogeneous Banks[†]

Marco Bellifemine
London School of Economics

Rustam Jamilov
All Souls College, University of Oxford

Tommaso Monacelli
Bocconi University, IGIER, CEPR

March 20, 2022

Abstract

We study how bank heterogeneity and market power shape the transmission of monetary policy. In the data, following a monetary contraction, large banks lower their *credit* mark-ups and quantities significantly more and raise their *deposit* mark-ups significantly less than small banks. We interpret our findings in a Heterogeneous Bank New Keynesian (HBANK) model, featuring permanent ("skill") and stochastic ("luck") bank returns heterogeneity, incomplete markets, variable asset and deposit market power, and nominal rigidities. In this setup, the aggregate effects of monetary policy shocks depend explicitly on the endogenous distribution of banks' net worth and the competitive structure of asset and deposit markets. Our main findings are three-fold. First, the model matches both the aggregate and the cross-sectional conditional responses to monetary policy shocks estimated from the data. Second, due to the *marginal propensity to lend* being increasing in profitability, permanent returns heterogeneity considerably strengthens the effectiveness of monetary policy shocks. Third, *credit* market power amplifies the effects of monetary policy shocks, whereas the opposite holds for *deposit* market power. This stems from credit mark-ups of large intermediaries being strongly conditionally pro-cyclical, while deposit mark-ups being counter-cyclical. When calibrated to U.S. bank-level data, the full model delivers substantial amplification of monetary shocks.

JEL codes: E44, E51, E52

Keywords: Financial intermediaries, heterogeneous agents, incomplete markets, monetary policy, market power

[†]Bellifemine: m.bellifemine@lse.ac.uk
Jamilov: rustam.jamilov@all-souls.ox.ac.uk
Monacelli: tommaso.monacelli@unibocconi.it

1 Introduction

The emphasis on the role of financial intermediaries in the transmission of monetary policy has strengthened after the 2007-08 financial and credit crisis. A large literature acknowledges that disruptions in financial intermediation can have significant effects on economic activity (Brunnermeier and Pedersen, 2009; Gertler and Kiyotaki, 2010; He and Krishnamurthy, 2013). However, most of the recent work in macroeconomics on the link between financial intermediation and monetary policy still abstracts from the implications of heterogeneity and imperfect insurance that characterize the full distribution of financial intermediaries. At the same time, a secular process of consolidation has reduced the competitive pressure in the banking industry (Corbae and D'Erasmo, 2020), renewing the attention for the role of competition and market power in shaping the pass-through of monetary policy to credit and deposit supply (Scharfstein and Sunderam, 2016; Drechsler et al., 2017; Corbae and Levine, 2018; Wang et al., 2020).

In this paper we integrate *bank heterogeneity* and *market power* to study the transmission of monetary policy in a Heterogeneous Bank New Keynesian (HBANK) model. Our framework combines five main features: (i) permanent ("skill") and stochastic ("luck") bank returns heterogeneity, (ii) time-varying asset and deposit mark-ups, (iii) endogenous bank default, (iv) deposit insurance, and (v) nominal rigidities. Our setup nests the canonical New Keynesian model (Woodford, 2003; Gali, 2008), the Gertler and Kiyotaki (2010) macro-banking framework, can accommodate heterogeneous or homogenous mark-ups in both credit or deposit markets, and features both permanent and idiosyncratic bank profitability heterogeneity.

In order to motivate and guide our quantitative analysis, we provide detailed micro-level evidence on financial (depository) institutions. We document a series of stylized facts on the *conditional* responses of key banking variables to monetary policy shocks identified with a high-frequency approach (Gurkaynak et al., 2005; Gertler and Kiyotaki, 2015). We highlight three main sets of facts, all in response to a monetary policy contraction. First, monetary policy activates a significant *banking market power* channel. On the asset side, on average, banks lower their mark-ups on loans, consistent with the goal of dampening the effects on the price of loans; on the liability side, however, banks increase their mark-ups on deposits, consistent with a so-called "deposit channel" of monetary policy (Drechsler et al., 2017).

Second, there is considerable cross-sectional variation in the observed response of banking market power. Large banks lower their *credit* mark-ups significantly more and increase their *deposit* mark-ups significantly less than small banks. Since in the cross section both credit and deposit mark-ups are increasing in size and profitability, the dispersion of mark-ups shrinks on the asset side while it widens on the liability side. This evidence points to a potential novel trade-off for monetary policy, between aggregate demand stabilization and efficiency in credit markets.

Third, assets, equity, leverage, and profitability all fall following a monetary policy contraction, pointing to a significant *average* bank balance sheet channel. There is, however, an important compositional effect, with large banks displaying a markedly higher elasticity of quantities (assets and net worth) in response to monetary shocks.

Taking stock of the above empirical facts, we require a micro-consistent framework that can deliver simultaneously the correct (i) *cross-sectional* relationships between market power, size, and profits, (ii) *conditional* responses of financial aggregates and mark-ups to transitory monetary policy shocks, and (iii) *heterogeneous* responses to monetary policy shocks by percentiles of the bank size distribution.

Our HBANK framework can rationalize all these facts. The centerpiece of our quantitative model are financial intermediaries (banks, for short) that feature market power on *both* the asset and liability sides of their balance sheet. They choose interest rates on deposits by setting mark-ups below the risk-free rate (i.e., mark-downs). The deposit mark-up arises endogenously because (i) deposits provide special liquidity services to the household and (ii) deposit products are imperfect substitutes with non-CES supply as in Kimball (1995). On the asset side banks also set credit prices, faced by non-financial firms, as a mark-up over the deposit rate. The variable credit mark-up arises due to non-CES demand for credit (Kimball, 1995). In addition to our departure from two-sided perfect competition, the second important feature of our model is market incompleteness and uninsured idiosyncratic bank returns risk in the spirit of Benhabib et al. (2019). We allow for *permanent* (ex-ante) and *stochastic* (ex-post) components of returns heterogeneity, both of which we recover directly from U.S. bank-level data. The interaction between returns heterogeneity and two-sided market power has powerful implications for the conduct of monetary policy.

We reach four main results. First, when calibrated to U.S. bank-level data, our model delivers considerable amplification of transitory monetary policy shocks. We find that much of that amplification comes from the presence of permanent returns heterogeneity. This stems from the fact that, in the model, the *marginal propensity to lend* (MPL), i.e., the elasticity of credit supply to banks' marginal cost shocks, is increasing in profitability; hence the presence of highly profitable intermediaries raises the aggregate credit supply elasticity of the economy.

Second, there is rich *heterogeneity* in the responses of banks in different percentiles of the size distribution. In line with the data, we find that in response to a monetary contraction large banks lower their quantities (assets) and credit mark-ups significantly *more* than small banks. Put differently, credit prices respond much less for large banks, consistent with a price pass-through being an inverse function of balance sheet size. On the other hand, on the liability side of banks' balance sheets, it is the *small* banks who increase their deposit mark-ups, and thus deposit rates, relatively more. In other words, heterogeneous, two-sided *real* rigidity in credit and deposit prices is a key feature of our environment, pointing to the importance of the interaction between bank

market power and balance sheet size as a driver of monetary policy transmission.

Third, credit market power amplifies the effects of monetary policy on both financial and macro variables. This stems from the conditional *pro-cyclicality* of credit mark-ups: credit mark-ups, particularly of large intermediaries, fall in response to a contractionary monetary policy shock. Intuitively, credit-side real rigidities make the loan price pass-through incomplete and bank assets more elastic. This is especially true for large banks, due to the elasticity of credit supply being increasing in size. On the other hand, deposit market power dampens the effects of monetary policy shocks, due to the *counter-cyclicality* of deposit mark-ups: deposit mark-ups rise in response to a contractionary monetary policy shock. This happens for two reasons. First, because of liability-side real rigidities, banks limit the pass-through of higher marginal funding costs onto the costs (i.e., remuneration) of deposits. Second, monetary contractions intensify households' preferences for deposit liquidity, which allows banks to effectively enjoy greater degrees of aggregate deposit market power endogenously. The two channels combined lead to a considerable increase in the weighted-average deposit mark-up (in other words, a large contraction in deposit mark-downs). Imperfect deposit rate pass-through implies that banks' balance sheet quantities decisions are exposed to a change in the cost of funds which is relatively smaller than under the benchmark perfect competition counterfactual. Both the (conditional) pro-cyclicality of credit mark-ups and the counter-cyclicality of deposit mark-ups are in line with our empirical results. We emphasize that these are *conditional* properties. Unconditionally, the asset-weighted credit mark-up is in fact counter-cyclical (when correlated against U.S. real GDP) while the deposit-weighted deposit mark-up is pro-cyclical. Properly conditioning on monetary innovations is thus important.

Fourth and finally, our model matches both unconditional and conditional moments of the banking data. For one, it is consistent with the positive cross-sectional relationship between bank size, profitability, and market power. That cross-section shows a long-term increase in the average degree of market power, on both sides of banks' balance sheets. In addition, the model matches several moments of the cross-sectional empirical response of banks' balance sheets and mark-ups to monetary policy shocks.

Literature review Our paper builds on several literature strands from macroeconomics, finance, and monetary economics. First, we contribute to a large literature that studies the aggregate consequences of bank market power. Existing studies analyze market power either on the asset or liability sides of bank balance sheet. Recent studies on credit market power include Corbae and D'Erasmo (2021) who build a quantitative macro-banking model with endogenous banking competition, Jamilov and Monacelli (2020) who study the role of monopolistic credit markets in a real business cycle model with counter-cyclical idiosyncratic bank return risk, and Wang et al. (2020) and Whited et al. (2021) who estimate credit mark-ups structurally, including in

the context of monetary policy transmission. Papers that quantify deposit market power include Drechsler et al. (2017, 2021) who establish the deposits channel of monetary policy transmission using branch-level data and Egan et al. (2017) who develop a structural model of the US banking sector featuring deposit-market competition and financial fragilities. Gerali et al. (2010) develop a dynamic stochastic general equilibrium framework (DSGE) model with both sticky credit and deposit rates, albeit in a representative-bank environment with complete markets. Relative to the literature, our contribution is that we allow for *endogenous and heterogeneous* credit *and* deposit mark-ups in a New Keynesian model with incomplete markets and in a framework that is consistent with detailed micro-level empirical analysis which we also provide.

Second, we are building on the fast-expanding literature on heterogeneous financial intermediaries. This literature can be further divided into two subsets. The first set studies situations where intermediaries feature *permanent* ex-ante heterogeneity. For example, Coimbra and Rey (2019) develop a general equilibrium framework with endogenous entry and where financial intermediaries are heterogeneous in their Value-at-Risk constraints. Begenau and Landvoigt (2022) build a quantitative model with two banking sectors that approximate the empirically-documented divide between standard commercial and "shadow" banks. The second subset of the literature introduces some form of bank-level non-systematic risk such that intermediaries are generally ex-ante identical but heterogeneous ex post. For example, Bianchi and Bigio (2022) study the credit channel of monetary policy in an environment where bank deposits circulate in an unpredictable way and banks face deposit withdrawal shocks. Rios Rull et al. (2020) study aggregate effects of capital requirements in a quantitative model with non-diversifiable credit risk. Relative to these two literature strands our contribution is to incorporate *both* stochastic and persistent bank returns heterogeneity, recover the two components directly from the data, and link them with monetary policy in a New Keynesian framework.

Third, our model builds on the so-called "macro-banking" literature which incorporates financial frictions into otherwise standard macroeconomic frameworks. There are two broad complementary directions in this literature. Some studies introduce *market*-based constraints on risk-taking that, generally speaking, generate counter-cyclical amplification of aggregate shocks. Papers in this strand include Gertler and Kiyotaki (2010), Gertler and Karadi (2011), Brunnermeier and Sannikov (2014), He and Krishnamurthy (2013), Jermann and Quadrini (2013), Nuno and Thomas (2016), Gertler et al. (2016), Gertler et al. (2020). On the other hand, studies such as Brunnermeier and Pedersen (2009), Adrian and Shin (2010), Adrian and Boyarchenko (2015) introduce book-based constraints on risk taking which thus do not differentiate between market or book leverage ratios. A notable exception is Begenau et al. (2021) who propose a unifying approach to modeling book vs market leverage of banks. Their empirically-motivated approach is based on delayed loss recognition that produces a wedge between book and "fundamental" balance sheet values. Our

paper adds to the market-based strand. However, our focus is on endogenous bank market power as a transmission mechanism for monetary policy, and not on the risk-taking channel. As such, in reduced form our market-based constraint provides a computationally convenient limit on leverage ratios without affecting our main results on credit or deposit mark-up cyclicality.

Finally, we are contributing to the vast literature that quantifies the role of heterogeneity, financial frictions, or both for monetary policy-making. Lee et al. (2020) introduce frictional financial intermediation into the canonical HANK model (Kaplan et al., 2018). Their representative-bank friction, which is similar to the one that we impose in our set-up, amplifies monetary policy and gives rise to consumption inequality. Our approach is different but conceptually similar: we focus on heterogeneous intermediaries but keep the household block very simple. Bigio and Sannikov (2021) build an incomplete-markets environment with wage rigidities where the central bank controls credit spreads and interest rate targets via the supply of reserves. In important related work, Baqaee et al. (2021) uncover the supply side of monetary policy in a model with heterogeneity and endogenous product market power of non-financial firms. Ottonello and Winberry (2020) quantify the investment channel of monetary policy in the case of non-financial firms that are heterogeneous in their riskiness and distance to default. Kaplan et al. (2020) emphasize the role of housing and long-term mortgages in the dynamic of credit conditions leading up to the Great Financial Crisis.[1] Our contribution is to zoom in both empirically and quantitatively on the heterogeneous *bank market power* channel of monetary transmission in an otherwise textbook New Keynesian model with endogenous capital accumulation and financial frictions.

The rest of the paper proceeds as follows. Section 2 details our empirical analysis. Section 3 develops a New Keynesian model with heterogeneous banks and endogenous credit and deposit market power. Section 4 describes how we bring the model to the data. Section 5 presents our quantitative results on the monetary transmission mechanism. Finally, Section 6 concludes.

2 Empirical analysis

In this section we document the behavior of credit and deposit mark-ups both over time and in the cross-section of banks. We also report the response of the distribution of several banking variables to identified monetary shocks. Section 2.1 presents the evolution of credit and deposit mark-ups over time as well as some stylized facts about their cross-sectional distribution. In Section 2.2 we look at the response of the distribution of the banking sector to identified monetary shocks. Section 2.3 presents additional results and robustness checks, as well as a discussion of how our

[1] As Haddad and Muir (2021) show, financial intermediaries are particularly important investors in credit and mortgage-backed securities markets. Introducing equilibrium house prices as drivers of bank market power, risk-taking, or both could be an important extension for future research.

measure of deposit mark-up relates to the deposit spread in Drechsler et al. (2017). Appendix A.3 shows the response of standard macroeconomic variables to monetary shocks under our estimation procedure. Finally, Appendices A.4 and A.5 describe more in detail, respectively, our estimation strategy for credit and deposit mark-ups and our data construction procedure.

2.1 Data description and variable definitions

Our main data source is the Federal Reserve Consolidated Reports of Condition and Income (also known as Call Reports). This dataset includes both income statement and balance sheet variables for the universe of U.S. FDIC-insured banks at a quarterly frequency. Our sample covers the period 1985q1-2020q1. The main variables of interest are assets, net worth (equity), net income, leverage, credit mark-ups, and deposit mark-ups. We extract assets, net worth, and net income directly from the data, while we construct book leverage as the ratio of assets over net worth.

Credit mark-ups To estimate *credit* mark-ups, we follow the procedure described in Corbae and D'Erasmo (2021) and proposed originally in De Loecker et al. (2020). In particular, we define the credit mark-up as the ratio of the price that banks charge on loans over the marginal cost of producing an extra unit of loans.[2]

$$\text{credit mark-up} = \frac{\textit{price of loans}}{\textit{marginal cost of loans}}$$

Crucially, the latter measure includes both interest and non-interest marginal costs, allowing us to retrieve a proper measure of mark-ups, as opposed to credit margins. We refer the reader to Appendix A.5 for a more detailed description of our methodology.

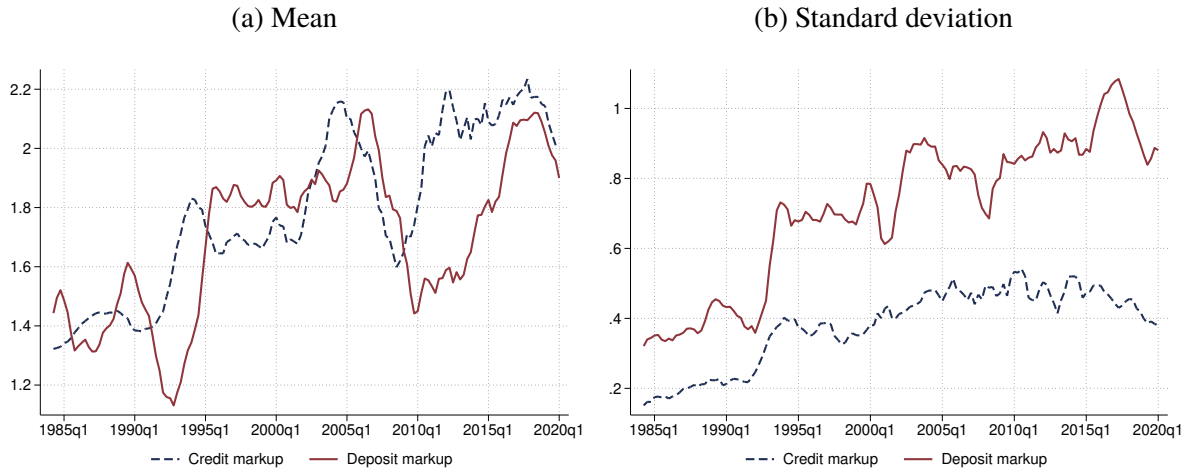
Deposit mark-ups We provide novel cross-sectional estimates of *deposit* mark-ups by applying to deposits a similar procedure employed in Corbae and D'Erasmo (2021) for credit mark-ups.[3] We define the deposit mark-up as the ratio of a proxy for the safe rate of return that banks are able to obtain out of their funds over the marginal cost of raising one additional unit of deposits.

$$\text{deposit mark-up} = \frac{\textit{safe return on funds}}{\textit{marginal cost of raising deposits}}$$

[2] See Appendix A.5 for more details on the estimation procedure.

[3] Wang et al. (2020) and Whited et al. (2021) estimate loan and deposit spreads to study, respectively, the transmission of monetary policy in a structural model where banks have market power and the interaction of that market power with risk-taking motives in a low-interest-rate environment.

Figure 1: Bank market power over time



Note: all series are smoothed using backward-looking moving averages of the last 4 observations. We weigh credit mark-ups by asset holdings and deposit mark-ups by deposit holdings.

In turn, the marginal cost of raising deposits is the sum of both the interest and the non-interest marginal cost of deposits. This allows us to derive a genuine measure of a deposit mark-up, rather than a deposit *spread* as, for example, in Drechsler et al. (2017). Once again, we refer the reader to Appendix A.5 for a more detailed description of our methodology.

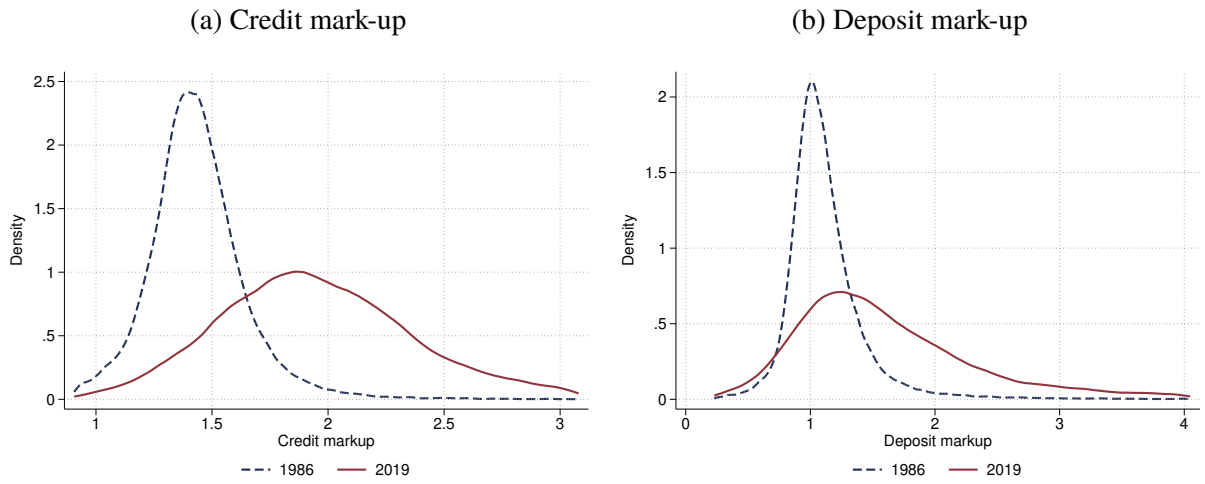
Figure 1 plots the evolution of the (weighted) average and standard deviation of the estimated credit and deposit mark-ups. We weigh credit mark-ups by asset holdings and deposit mark-ups by deposit holdings. As already documented in Corbae and D'Erasmo (2021), we find an increasing trend in the average credit mark-up.[4] At the same time, to the best of our knowledge, we are the first to document an upward trend in the average deposit mark-up. Note that this finding is not in contrast with the downward trend in the deposit *spread* highlighted in Drechsler et al. (2017). In fact, the weighted average of the deposit spread has been trending downward over time in our sample as well.[5] We also document a sharp upward trend in the *dispersion* of both credit and deposit mark-ups over time, in line with what has been highlighted elsewhere by De Loecker and Eeckhout (2021) for the whole U.S. economy. This marked and steady rise in the dispersion of mark-ups points to a potential increased inefficiency both in the credit and the deposit markets, and calls for a closer look at the evolution of the whole distributions of mark-ups.

Figure 2 plots the estimated densities of credit and deposit mark-ups in years 1986 and 2019. As already hinted above, there is a stark increase in the dispersion of both variables. Moreover,

[4]Notice that our time series for the weighted average of credit mark-ups is in line with the estimates in Corbae and D'Erasmo (2021) as well as those in De Loecker et al. (2020) for the "Finance and Insurance" industry.

[5]See Section 2.3 for a more detailed discussion of the relationship betweeen the deposit spread and our estimated deposit mark-up.

Figure 2: Distribution of bank market power - 1986 and 2019



the increased mass in the right tails of the distributions point to a marked rise in the concentration of both credit and deposit mark-ups. Note that a very similar trend is observed in De Loecker and Eeckhout (2021) for the case of non-financial firms.

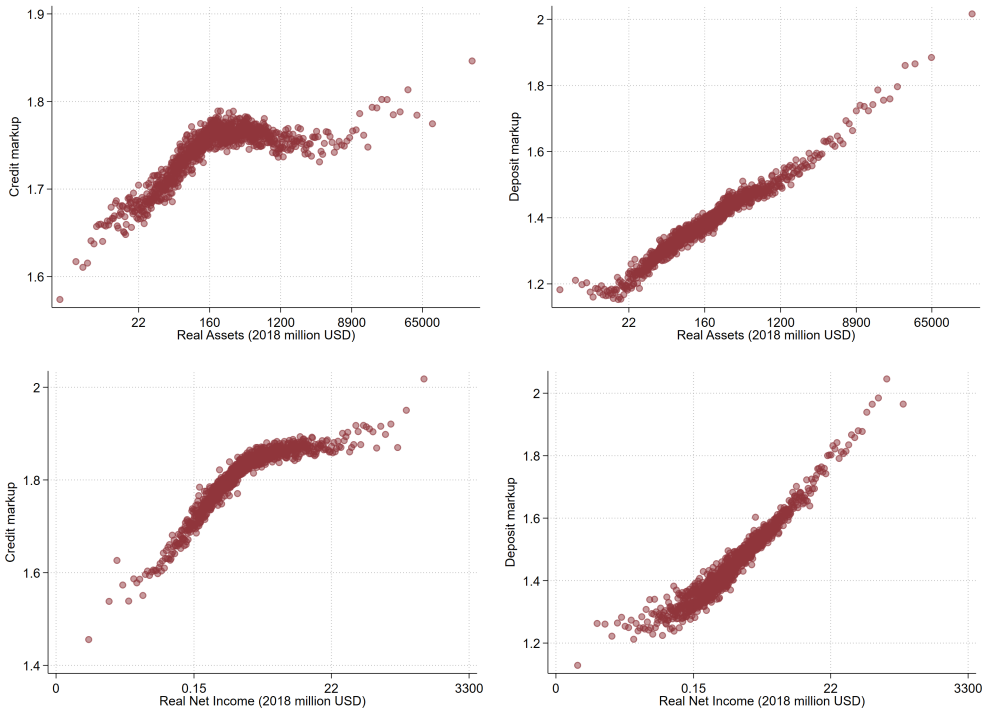
Taken together, both the upward trend in average mark-ups and the simultaneous increase in dispersion and concentration, hint at the importance of *time-varying heterogeneity* when measuring credit and deposit market power.

Mark-ups, size and profitability Next we look at how credit and deposit market power correlate with other observable dimensions of bank heterogeneity.

The top panels of Figure 3 show binned scatter plots of bank *size* –defined as total book assets– against credit and deposit mark-ups. We first divide our sample into 1,000 equally sized bins of log assets, each including roughly 1,250 observations. We then residualize both x-axis and y-axis variables by a time fixed effect. Finally, for each bin we plot the unweighted average of the x-axis variable, as well as the weighted average of our mark-up variable. We find that both credit and deposit mark-ups are positively correlated with bank size. That is, larger banks charge a larger mark-up over their marginal cost of producing one extra unit of credit, and they also charge a larger mark-up over their marginal revenue from one extra unit of deposits. The bottom panels of Figure 3 perform a similar exercise, by looking at the relation between mark-ups –both on the credit and the deposit market– and net income, which we use as a proxy for profits. We show that, after controlling for time fixed effects and bank size, more profitable banks display larger credit and deposit mark-ups.[6] While Corbae and D'Erasmo (2021) and Benetton (2021) already highlight

[6]Notice that by taking the logarithm of the variable on the x-axis we discard observations with negative net income. Figure A.11 in Appendix A.2 shows that the positive relationship between mark-ups and profits holds even when we proxy profitability by return on equity (RoE) instead of net income.

Figure 3: Bank size, market power, and profitability



Note: the x-axis is in logs. We control for time fixed effects in the top panels, and for time fixed effects as well as assets in the bottom panels.

the relationship between credit mark-ups and size, a contribution of our work is to document two novel facts: (i) the positive correlation between deposit mark-ups and bank size; (ii) the positive correlation between profits and (credit and deposit) mark-ups.

Summary of main facts To summarize, the unconditional descriptive analysis highlights two main facts. First, there is an upward trend in the average, dispersion, and concentration of bank credit and deposit market power. Second, there is a positive correlation between (credit and deposit) mark-ups and bank size and profits. These two facts, combined with the vastly documented rising trend in banking concentration, suggest that a fat tail of big, profitable banks with large credit and deposit market power has become increasingly relevant for the banking sector. We now turn to investigate whether this matters for monetary policy and, more specifically, to what extent monetary policy interacts with different dimensions of bank heterogeneity.

2.2 Responses to monetary policy shocks

We present empirical evidence on the response of the distribution of selected banking variables to identified monetary policy shocks. We estimate a proxy-SVAR model as proposed in Stock

and Watson (2012) and Mertens and Ravn (2013), and used in Gertler and Karadi (2015). We instrument monetary shocks with the change in the 3-month ahead Fed Funds futures within 30 minutes windows around FOMC announcements.[7] Baseline estimation of the VAR model, from which we obtain the reduced form residuals, is based on the monthly sample 1985:01-2017:12. Since our series of high-frequency shocks is available only starting from 1990:02, we use the IV strategy over the restricted sample 1990:02-2017:12 only. The baseline VAR specification reads:

$$X_t = \mu + \sum_{j=1}^{12} A_j X_{t-j} + u_t \tag{1}$$

where μ is a constant, u_t is the vector of reduced form residuals –which will be projected onto the monetary policy shock instrument– and X_t is a vector that includes (i) Fed Funds rate, (ii) consumer price index, (iii) industrial production index, (iv) S&P 500 index monthly return, and (v) the moment of a selected banking variable.[8]

Figure 4 plots the aggregate response of selected banking variables to a one standard deviation *contractionary* monetary policy shock. The figure displays the response of total assets, net worth, leverage, and net income; asset-weighted averages for credit mark-ups; and deposit-weighted averages for deposit mark-ups.[9] The average size of the banking sector, as proxied by either book asset holdings or net worth, declines by around 0.5%. We also document a substantial decrease in net income in response to the shock. While this latter finding may look at odds with the "conventional wisdom" that higher interest rates are beneficial for banks' profitability, research on the topic has reached contrasting results.[10] Leverage, instead, does not show any significant response to monetary shocks, in line with what Miranda-Agrippino and Rey (2020) find for commercial U.S. banks. As for mark-ups, we document that after a contractionary monetary shock, the average credit mark-up decreases by roughly 1%, while the average deposit mark-up increases by a comparable, but slightly smaller, magnitude. These results suggest that, through movements in the mark-up, bank market power may play an important role for the transmission of monetary policy, as it generates an incomplete pass-through of monetary shocks into both loan and deposit rates.[11]

[7] In Appendix A.2 we show that our results are unaffected if we use instruments for monetary shocks that control for the information content of FED's announcements, as in Jarociński and Karadi (2020).

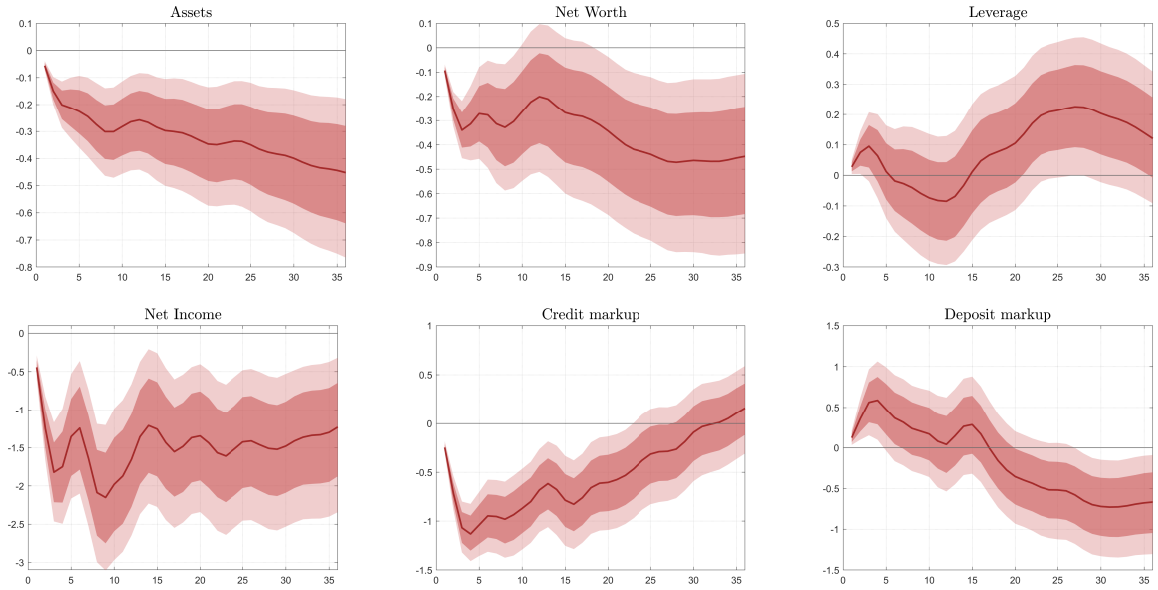
[8] Note that, as explained more in detail in Appendix A.4, we interpolate banking variables from the original quarterly frequency to monthly frequency. In Appendix A.2 we perform a variety of robustness checks, where we show that the bulk of our results is unchanged if we use variables at the original quarterly frequency or if we look at different sample windows.

[9] We weigh banks based on their average asset or deposit holdings over the last 4 quarters. See Appendix A.4 for more details.

[10] Altavilla et al. (2018), for example, find that the return on assets of European banks does not deteriorate following monetary expansions.

[11] Incomplete pass-through of monetary shocks to loan and deposit rates has also been documented in Wang et al. (2020), Drechsler et al. (2017) and Polo (2021).

Figure 4: Aggregate responses to a contractionay monetary policy shock



Notes. Lightly shaded areas represent 90% wild bootstrap confidence intervals based on 10,000 draws. Darkly shaded areas are 68% confidence intervals. The y-axis is in percentage points, while the x-axis represents months elapsed since the shock. Assets, net worth, and net income are within-period totals; leverage is total assets over total net worth; we use asset weighted average for credit mark-ups and deposit weighted average for deposit mark-ups. See Appendix A.4 for details on the weighting scheme.

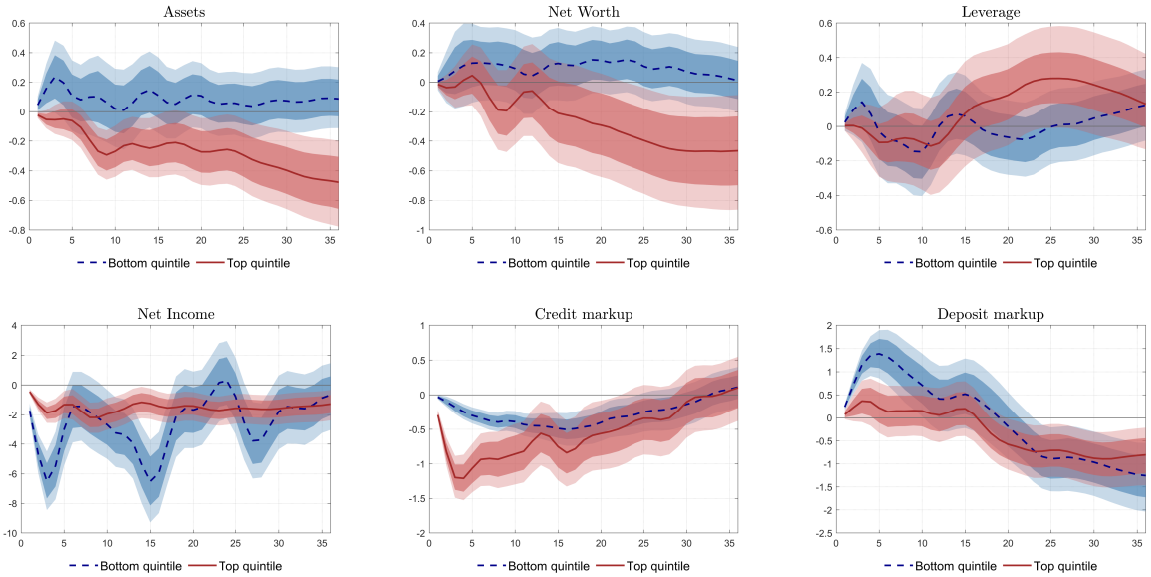
Monetary shocks in the banking cross-section Our previous results highlighted wide heterogeneity in the distribution of credit and deposit mark-ups in the cross-section of banks and showed that a big chunk of this heterogeneity develops along the size distribution. We now turn to the differential response of banking variables to monetary shocks across the size distribution. For each period in our sample we split observations into *size quintiles* based on banks' average asset (deposit) holdings over the last four quarters, in line with the weighting scheme previously used for net income and mark-ups.[12] We then compute an aggregate measure of the banking variables of interest within each quintile, following the same procedure used in Figure 4. Finally, we estimate the response of the banking variables to a monetary shock across different size quintiles by using the same VAR specification as above.

Figure 5 shows the results. Three main facts are worth emphasizing. First, large (top-quintile) banks reduce their asset holdings relatively more than small (bottom-quintile) banks. In addition, banks in the top 20% of the asset distribution also shrink their net worth by more in response to a monetary tightening.[13] Second, small banks experience a larger contraction in net income relatively

[12]See Appendix A.4 for more details. Notice that our results are robust to different size classifications and do not change if we look, for example, at assets deciles.

[13]The aggregate size of banks in the bottom quintile of assets, as proxied by either assets or net worth, actually shows a slight increase after the shock, even though the response is never significantly different from zero.

Figure 5: Cross-sectional responses to a contractionary monetary policy shock



Notes. Assets, net worth, leverage, and net income are totals within the quintile. Credit mark-ups are asset-weighted within the quintile. Deposit mark-ups are deposit-weighted within the quintile. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

to large banks. The response of leverage, instead, does not display any clear variation across size. Third, the elasticity of credit mark-ups to monetary shocks is substantially larger for large banks relatively to small ones. Coupled with the finding that large banks have larger credit mark-ups, this evidence suggests that the bulk of the incomplete pass-through of monetary shocks to credit rates is driven by big banks absorbing a larger share of the interest rate increase. On the other hand, the opposite is true for deposits. Relative to large banks, small banks display a significantly larger response of deposit mark-ups to monetary policy shocks.

2.3 Mark-ups, spreads and robustness checks

In Section 2.1 we have documented an upward trend in the average deposit mark-up. However Drechsler et al. (2017) highlight a downward trend in the deposit *spread*. These two facts are not necessarily at odds with each other. In fact, the deposit spread is defined as the logarithm of the ratio of a proxy of the safe rate of return over the deposit rate. Our definition of the deposit mark-up relies on the same ratio, but it includes marginal net non-interest expenses of deposits in the denominator. As a result, the two measures do not need to show perfect co-movement, and a time-varying wedge can appear between them. Figure A.13 presents a decomposition of our measure of deposit mark-up into its primitive components. In line with Drechsler et al. (2017), we also find a decrease in the average deposit spread over time. However, the decreasing trend in

marginal non-interest expenses generates an upward trend in the average deposit mark-up. We thus document a decoupling between the deposit spread and the deposit mark-up, driven by a downward trend in banks' marginal *non-interest* expenses.

Figures A.3 to A.10 in Appendix A.2 show that our VAR results are robust to a variety of specifications. In particular, all our findings are unaffected if we employ the Jarociński and Karadi (2020) series for monetary shocks, that controls for the information content of Fed's announcements. Similarly, starting our sample in 1990:01 or ending it in 2012:06 has no effect on the results. Moreover, the bulk of the aggregate and heterogeneous responses is qualitatively the same if we consider quarterly, rather than monthly, data.

3 A New-Keynesian model with heterogeneous banks

To shed light on the empirical facts emphasized above, we build a New Keynesian model with heterogeneous financial intermediaries, uninsured idiosyncratic rate of return risk, and endogenous (asset and deposit) mark-ups. The economy is populated by a continuum of monopolistically competitive banks indexed by $j \in [0, 1]$, a representative household, producers of capital, intermediate, and final goods, and a monetary policy authority. There is no aggregate uncertainty.

3.1 Households and deposit markets

Time is discrete and infinite. The representative household supplies labor inelastically (normalized to unity) and derives utility from consumption and liquid wealth. The household can save in the form of one-period deposits or mutual funds. Deposits provide special liquidity services, similarly to the set-up of Drechsler et al. (2017, 2021) or more generally to the money-in-utility framework (Sidrauski, 1967; Gali, 2008; Walsh, 2010). Mutual funds are risk-less investments but provide no liquidity utility. Both vehicles pay guaranteed, state non-contingent rates of returns.

The utility index, which is increasing and strictly concave in consumption and deposit holdings, is defined as:

$$U(C_t, B_t) = \frac{C_t^{1-\psi}}{1-\psi} + \frac{B_t^{1-\nu}}{1-\nu} \tag{2}$$

where $\frac{1}{\nu}$ is the elasticity of deposit supply and $\frac{1}{\psi}$ is the intertemporal elasticity of substitution. Deposit products are imperfect substitutes across banking franchises, indexed by j. The deposit market is monopolistically competitive and aggregate deposit supply is given, as in Kimball (1995), by the aggregator:

$$\int_0^1 \Upsilon\left(\frac{b_{j,t}}{B_t}\right) dj = 1 \tag{3}$$

where $\Upsilon(x)$ is a strictly increasing and convex function. The consumer maximizes utility subject to the budget constraint:

$$C_t + \int_0^1 b_{j,t}dj + M_t \leq R_t M_{t-1} + \int_0^1 R^b_{j,t} b_{j,t-1} + W_t + Div_t + T_t \tag{4}$$

where M_t are mutual fund holdings, W_t is the competitive wage rate, $R^b_{j,t}$ is the non-contingent bank-specific interest rate on deposits to be determined in equilibrium, R_t is the real risk-free interest rate, T_t are lump-sum taxes, and Div_t are periodic lump-sum transfers of bank dividends. The first-order condition with respect to deposit supply yields:

$$R^b_{j,t+1} = R_{t+1}\left(1 - \underbrace{\left[\frac{C_t^{\psi}}{B_t^{\nu}} \frac{\Upsilon'\left(\frac{b_{j,t}}{B_t}\right)}{\mathcal{A}^b_t}\right]}_{\text{deposit spread} \geq 0}\right) \tag{5}$$

where $R_{t+1} = \left[\beta \mathbb{E}_t \left(\frac{C_{t+1}}{C_t}\right)^{-\psi}\right]^{-1}$ is the risk-free rate, determined via a first-order condition with respect to mutual fund holdings, and $\mathcal{A}^b_t := \int_0^1 \Upsilon'\left(\frac{b_{j,t}}{B_t}\right) \frac{b_{j,t}}{B_t} dj$. Derivations are shown in Appendix B.1. Notice that, since $\Upsilon(.)$ is convex, the deposit spread is positive and increasing in the relative size $\frac{b_{j,t}}{B_t}$. This means that deposit market power is concentrated in the right tail of the bank size distribution.

Deposit mark-up The implied deposit mark-up is defined as:

$$\mu^b_{j,t} = \frac{R_{t+1}}{R^b_{j,t+1}} \geq 1 \tag{6}$$

The above flexible specification can nest multiple environments. First, in the baseline scenario, a finite ν coupled with Kimball aggregation deliver endogenous and heterogeneous deposit mark-ups through two channels: (i) the real rigidity channel; (ii) the liquidity preference channel. Second, in the limit of $\nu \to \infty$, liquidity preferences disappear and the deposit spread drops to zero for all banks. Third, in the special case of a CES deposit aggregator, the deposit mark-up is homogenous across banks and is proportional to a constant elasticity of substitution. Finally, when deposit products are perfect substitutes and real rigidities disappear completely, the (homogenous) mark-up is rationalized only by deposits' special liquidity services.

3.2 Capital production and asset markets

Capital is required for the production of a final good. Capital good producers are cash-strapped and require bank financing in the form of equity-type claims. We assume that these firms possess a technology to costlessly convert claims into differentiated units of capital, which get immediately aggregated. The asset market is monopolistically competitive and aggregate capital K_t is assembled according to the Kimball-type aggregator:

$$\int_0^1 \Phi\left(\frac{k_{j,t}}{K_t}\right) dj = 1 \tag{7}$$

where $\Phi(x)$ is a strictly increasing and concave function. Firms solve the following problem:

$$\max_{k_{j,t}} \left[Q_t K_t - \int_0^1 q_{j,t} k_{j,t} dj \right]$$

subject to 7.

The solution to the above problem yields the following inverse asset demand curve:

$$\frac{q_{j,t}}{Q_t} \mathcal{A}_t^k = \Phi'\left(\frac{k_{j,t}}{K_t}\right) \tag{8}$$

where $\mathcal{A}_t^k := \int_0^1 \Phi'\left(\frac{k_{j,t}}{K_t}\right) \frac{k_{j,t}}{K_t} dj$, and $Q_t = \int_0^1 q_{j,t} \frac{k_{j,t}}{K_t} dj$ is the aggregate price index. Capital depreciates fully every period. A Lerner-style decomposition of q_j into mark-ups and marginal costs is shown in the next section.

3.3 Banks

Balance sheet Banks intermediate funds between households and capital producing firms. In order to motivate an invariant dividend payout rule, we assume that every period a fraction of banks $1 - \sigma$ exits the market exogenously. Endogenous exit is also allowed and described further below. Banks start the period with some initial net worth n_j and must choose firm claims k_j, deposit demand d_j, the price of claims q_j, and the deposit rate R_j^b, subject to the balance sheet constraint:

$$d_{j,t} + n_{j,t} = k_{j,t} \tag{9}$$

The law of motion of bank net worth is given by:

$$n_{j,t+1} = R_{j,t+1}^T q_{j,t} k_{j,t} - R_{j,t+1}^b d_{j,t} - \zeta_1 k_{j,t}^{\zeta_2} \tag{10}$$

where $R^T_{j,t+1}$ is bank j's *total* return on assets, and ζ_1 and ζ_2 are parameters that govern non-interest expenses (or, alternatively, asset adjustment costs). The parameter ζ_1 is useful for calibrating the average size of the balance sheet in the steady state while ζ_2 is important for delivering non-linearity and scale-variance.

Return heterogeneity We assume that financial markets are incomplete and banks earn bank-specific returns that consist of a permanent and a stochastic component:

$$R^T_{j,t} = \underbrace{\kappa_j}_{\text{permanent}} + \underbrace{\rho_\xi \xi_{j,t-1} + \sigma_\xi \epsilon_{j,t}}_{\text{stochastic}} \tag{11}$$

Permanent heterogeneity in returns κ_j captures the fact that some banks are intrinsically more skilled than others in terms of identifying profitable lending opportunities. Time varying, stochastic heterogeneity in returns $\xi_{j,t}$ implies that exposure to firm-specific risk cannot be perfectly hedged away with derivatives. This assumption is motivated by a growing literature that argues that idiosyncratic (non-aggregate) risk matters for bank outcomes (Amiti and Weinstein, 2018; Galaasen et al., 2021). Later in the paper we will recover κ and ξ directly from U.S. bank-level data.

Leverage constraint We allow for an occasionally binding equity-based constraint on bank leverage.[14] Following Gertler and Karadi (2011) and Gertler and Kiyotaki (2010), the bank-household relationship features a moral hazard friction. Banks have an incentive to divert franchise assets and have the ability to divert no more than a fraction λ of total market value of assets $q_{j,t}k_{j,t}$. Conditional on diverting, the banker always escapes, but the franchise enters bankruptcy the following period. The banker is indifferent between operating honestly and diverting when the amount she is able to finance exactly equals the value of the franchise. This yields the following incentive constraint:

$$\lambda q_{j,t} k_{j,t} \leq V_{j,t} \tag{12}$$

What is different in our model relative to the representative-intermediary case is that the Lagrange multiplier on the leverage constraint is *bank-specific*, i.e., the constraint may bind for some banks while remaining slack for others, thus generating rich equilibrium non-linearities.

Default risk and deposit insurance Each bank can default with its own endogenous probability s_j. Default risk is due to fundamental insolvency, i.e., when bank net worth at normal market prices

[14] All our main quantitative results on the bank market power channel of monetary transmission remain unchanged if we adopt a debt-based leverage constraint in the spirit of Brunnermeier and Pedersen (2009) or Adrian and Shin (2010) and Adrian and Boyarchenko (2015). Alternatively, one may adopt the unifying approach in Begenau et al. (2021) and focus on the dynamics of both book and market leverage simultaneously.

is non-positive:[15]

$$s_{j,t} = \Pr\left(n_{j,t+1} \leq 0\right) \tag{13}$$

Conditional on bank j's insolvency, the household recovers only a fraction of promised payments $x_{j,t} \leq 1$, the risk that is priced by the household into deposit rates. Remaining assets get transferred to the capital producing firm who produces K_t as normal. Retail and final good production then resume as before.

Deposits are risky investments since bank default risk is priced competitively into the deposit rate distribution. In practice, however, banking sectors worldwide feature deposit insurance mechanisms which insulate households (up to a limit) from this risk. This constitutes one of the essential pillars of financial intermediation (Farhi and Tirole, 2021). We therefore introduce a parsimonious deposit insurance friction. Banks and households ex ante perceive that $s_j = 0 \; \forall \; j$ while optimizing. However, each bank can still default ex post depending on the actual realization of the idiosyncratic shocks ξ. In other words, there is a disconnect between fundamental insolvency risk and deposit market pricing. For parsimony, we assume that default is costless such that the only distortion that deposit insurance generates is through prices. The scheme is financed with lump-sum non-distortionary taxation of the household.

Dynamic problem We are now ready to describe the full dynamic optimization problem of the banking sector. We drop the subscript j notation temporarily. The state vector consists of net worth n, permanent profitability type κ, and idiosyncratic return draw ξ. Banks choose the size of the balance sheet and prices in both credit and deposit markets:

$$\max_{\{k_t, q_t, d_t, R_t^b\}} V_t(n_t, \kappa_t, \xi_t) = \mathbb{E}_t\left[\Lambda_{t+1}\left((1-\sigma)n_{t+1} + \sigma V_{t+1}\right)\right] \tag{14}$$

subject to:

$$\begin{aligned}
n_{t+1} &= R_{t+1}^T q_t k_t - R_{t+1}^b d_t - \zeta_1 k_t^{\zeta_2} \\
d_t + n_t &= k_t \\
R_{t+1}^T &= \kappa + \xi_{t+1} \\
\lambda q_t k_t &\leq V_t \\
R_t^b &= R_t\left(1 - \left[\frac{C_t^{\psi}}{B_t^{\nu}} \frac{\Upsilon'\left(\frac{b_t}{B_t}\right)}{\mathcal{A}_t^b}\right]\right) \\
\frac{q_t}{Q_t}\mathcal{A}_t^k &= \Phi'\left(\frac{k_t}{K_t}\right)
\end{aligned}$$

[15] We abstract from financial panics (bank runs) in this paper. Recent work on equilibrium illiquidity crises includes Uhlig (2010) and Gertler et al. (2020).

Note that banks internalize both the monopolistic credit demand and deposit supply systems.[16]

Credit mark-up and marginal cost The banking problem above can be shown to yield a Lerner-type condition that decomposes the price of claims into the credit mark-up and the marginal cost:

$$q_{j,t} = \underbrace{\frac{\sigma_{j,t}}{\sigma_{j,t}-1}}_{\text{credit mark-up } \mu^{k}_{j,t}} \underbrace{\frac{R^{b}_{j,t} + \zeta_1 \zeta_2 k_j^{\zeta_2 - 1}}{R^{T}_{j,t+1}}}_{\text{marginal cost } mc_{j,t}} \tag{15}$$

The marginal cost has three components: (i) the cost of external financing R^{b}_{j}, (ii) the non-interest cost of balance sheet expansion, and (iii) the heterogeneous return R^{T}_{j}, the latter acting essentially as a cross-sectional "productivity shifter". The credit mark-up depends on the (endogenous and heterogeneous) credit demand elasticity σ_j, which is determined by the particular parameterization of the Kimball demand system.

Heterogeneity in the marginal propensity to lend We now define an object that is important for bridging microeconomic heterogeneity in the credit market and macroeconomic fluctuations: the *marginal propensity to lend* (MPL). MPL measures the response of credit supply to a change in the marginal cost: $MPL_j := \frac{\partial k_j}{\partial mc_j}$. We are particularly interested in computing the semi-elasticity of credit supply with respect to gross returns:

$$\frac{\partial k_j}{\partial \log R^{T}_{j}} = \underbrace{-\sigma_j \frac{k_j}{mc_j} \frac{1}{1+\Omega_j}}_{MPL_j} \left(-\frac{1}{R^{T}_{j}} \right) > 0 \tag{16}$$

where Ω_j is the elasticity of the credit mark-up to marginal cost shocks for bank j. The first object in the above equation, the demand elasticity $-\sigma_j$, is strictly negative. The second object is strictly positive, Ω_j being the price pass-through, which is between zero and unity. The final term is the semi-elasticity of the marginal cost with respect to returns, which is negative. Thus, credit supply, in partial equilibrium, increases with the returns profile R^{T}_{j}. Although stochastic return risk ξ_j may average out over time, the permanent component κ_j matters: persistently more profitable banks are effectively more "productive" and have a consistently greater lending elasticity. This, in turn, allows them to accumulate a significant amount of net worth and outgrow the leverage

[16] They do not, however, internalize the impact of their private choices on *aggregate* quantities and prices. This leads to a type of aggregate demand externality (Blanchard and Kiyotaki, 1987; Farhi and Werning, 2016). We abstract from normative implications in this paper.

constraint. As the size of the balance sheet expands, the endogenous competition structure leads to a change in aggregate credit and deposit mark-ups, because both are not size invariant due to Kimball aggregation. Omission of κ_j could therefore drastically underestimate the role of bank heterogeneity and market power for monetary policy transmission.

Specifying deposit and asset market structures We adopt the Klenow and Willis (2016) parametric specification for both asset and deposit markets. Credit and deposit market aggregators are defined accordingly:

$$\Phi\left(\frac{k}{K}\right) = 1 + (\theta_k - 1)\exp\left(\frac{1}{\epsilon_k}\right)\epsilon_k^{\frac{\theta_k}{\epsilon_k}-1}\left[\Gamma\left(\frac{\theta_k}{\epsilon_k}, \frac{1}{\epsilon_k}\right) + \Gamma\left(\frac{\theta_k}{\epsilon_k}, \frac{\left(\frac{k}{K}\right)^{\epsilon_k/\theta_k}}{\epsilon_k}\right)\right] \tag{17}$$

$$\Upsilon\left(\frac{b}{B}\right) = 1 + (\theta_b - 1)\exp\left(\frac{1}{\epsilon_b}\right)\epsilon_b^{\frac{\theta_b}{\epsilon_b}-1}\left[\Gamma\left(\frac{\theta_b}{\epsilon_b}, \frac{1}{\epsilon_b}\right) + \Gamma\left(\frac{\theta_b}{\epsilon_b}, \frac{-\left(\frac{b}{B}\right)^{\epsilon_b/\theta_b}}{\epsilon_b}\right)\right] \tag{18}$$

where $\Gamma(.,.)$ is the incomplete Gamma function. Parameters $\{\theta_k, \theta_b\}$ help control the average credit and deposit mark-ups. Parameters $\{\epsilon_k, \epsilon_b\}$ help determine the slopes of the credit and deposit mark-up functions. Note that the resulting relative credit demand and deposit supply curves are downward and upward sloped, respectively. We list additional formulae, such as derivatives of $\{\Phi, \Upsilon\}$ in Appendix B.2.

3.4 Non-Financial firms

Non-financial firms consist of a final good producer and of a continuum of differentiated retailers, indexed by $i \in [0, 1]$, that produce intermediate goods.

Final good production. Differentiated goods produced by retailers are aggregated into the final good by the final good producer:

$$Y_t = \left(\int_0^1 y_{i,t}^{\frac{\gamma-1}{\gamma}} di\right)^{\frac{\gamma}{\gamma-1}} \tag{19}$$

where γ>1 is the elasticity of substitution between differentiated goods.

Intermediate goods production. Each retailer rents labour competitively and buys the total capital stock to produce intermediate goods using a constant returns to scale production technology.

$$y_{i,t} = A_t K_{i,t}^{\alpha} L_{i,t}^{1-\alpha} \tag{20}$$

Retailers set a relative price for their variety $p_{i,t}$ and pay quadratic adjustment costs $\frac{\varphi}{2}\left(\frac{p_{i,t}}{p_{i,t-1}}-1\right)^2 Y_t$. The demand function for each retailer is: $y_{i,t} = \left(\frac{p_{i,t}}{P_t}\right)^{-\gamma} Y_t$ where $P_t = \left(\int_0^1 p_{i,t}^{1-\gamma} di\right)^{\frac{1}{1-\gamma}}$ is the relative price index. Cost minimization yields the following expression for the (common) nominal marginal cost: $MC_t = \frac{1}{A_t}\left(\frac{w_t}{1-\alpha}\right)^{1-\alpha}\left(\frac{Z_t}{\alpha}\right)^{\alpha}$, where Z_t is the rental cost of capital. The final good gets consumed every period.

Phillips curve Retailers' symmetrical problem yields the conventional Phillips curve relationship:

$$\log \Pi_t = \frac{\gamma-1}{\varphi}(\log MC_t - \log MC^*) + E_t\left[\Lambda_{t+1}\log \Pi_{t+1}\right] \tag{21}$$

where Π is gross inflation. The Phillips curve links heterogeneous bankers with the New Keynesian block. If demand for the final good Y_t is high, retailers increase production of their goods because of nominal rigidities. Demand for the differentiated good $y_{j,t}$ increases and the relative price P_t increases. Inflation goes up. Higher inflation in turn reduces the real rate of return in the economy and lowers the bankers' discount rate, spurring demand for deposits, higher leverage, and capital good production. Because of heterogeneity in the marginal propensity to lend, total expansion in credit and production will depend on the distribution of bank net worth. Additionally, because of two-sided market power, individual banks adjust both credit and deposit mark-ups. Conditional on the degree of credit and deposit price pass-through, this leads to a second-round effect on aggregate production of capital, consumption, and household's demand for bank deposits. Deposits' special liquidity status, being fully internalized by individual banks, further affects leverage-taking and credit supply, and so on.

In the next sections, we will provide a full decomposition of the total response of aggregate output to a monetary policy shock into partial and general equilibrium components.

3.5 Monetary policy

The monetary policy authority sets the nominal interest rate according to the Taylor-type rule:

$$i_t = \bar{R} + \phi_\pi \Pi_t + \epsilon_{m,t} \tag{22}$$

with ϕ_π is the weight on inflation and $\epsilon_{m,t}$ is a random disturbance drawn from a Normal distribution $\mathcal{N}(0, \sigma_m)$.

3.6 Equilibrium

Let $\mathbf{s} = \{n, \kappa, \xi\}$ and $\mathbf{S} = \{\Lambda, K, B, \Pi\}$ be the (bankers') idiosyncratic and aggregate state vectors, respectively. A stationary equilibrium for this economy is given by a set of aggregate $\{w, R\}$ and bank-level prices $\{q_j, R^b_j\}$ such that (a) given these prices policy functions of the bankers and households solve their respective decision problems, (b) aggregates are consistent with stationary distributions, and (c) all markets clear. We solve the model non-linearly. The main quantitative exercise involves computing transitional dynamics in response to unexpected "MIT" monetary policy shocks. Computation of the deterministic transition path is based around the shooting algorithm detailed in Boppart et al. (2018). Description of our algorithms is provided in Appendix B.4. This concludes the description of the model.

4 Taking the model to the data

We parameterize our model in several steps. First, we discuss how we recover permanent and stochastic returns heterogeneity directly from U.S. bank-level data. Second, we calibrate $\{\theta_k, \theta_b\}$ in order to target the average credit and deposit mark-ups that we estimate in Section 2. Finally, we rely on prior literature to fix the remaining parameters exogenously.

Returns heterogeneity We want to empirically decompose bank returns $R^T_{j,t}$ into the bank fixed effect (permanent heterogeneity) κ_j, and the AR(1) error term (stochastic heterogeneity) $\xi_{j,t}$:

$$R^T_{j,t} = \kappa_j + \xi_{j,t} \tag{23}$$

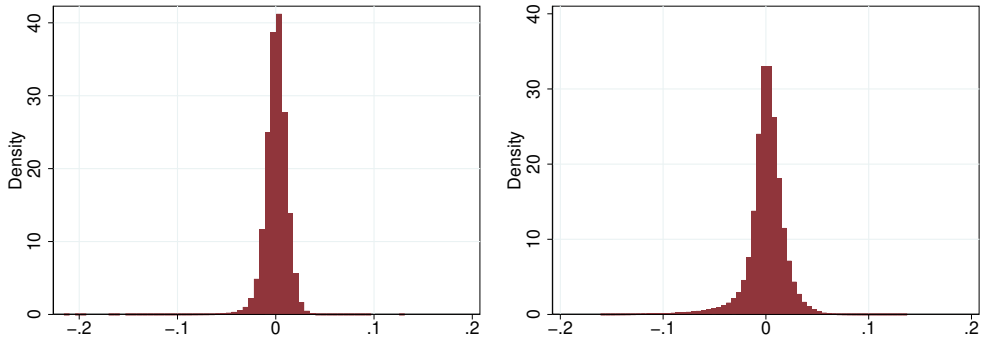
where

$$\xi_{j,t} = \rho_\xi \xi_{j,t-1} + \sigma_\xi \epsilon_{j,t} \tag{24}$$

where $|\rho_\xi| < 1$ and $\epsilon_{j,t}$ is an i.i.d random disturbance. Our data source for this exercise is the U.S. Call Reports, i.e. the same as in Section 2. Our baseline definition of returns is returns on equity or ratio of net income to total equity. We compute total equity as the difference between total assets and total liabilities. Our quarterly final sample runs from 1976q1 until 2020q1. We proceed in two basic steps. First, we de-mean the returns series by deducting the quarterly average of returns from $R^T_{j,t}$. Second, on the demeaned sequence, we run the Baltagi and Wu (1999) linear panel fixed effects model with AR(1) disturbances. We employ the Durbin-Watson estimator for ρ_ξ.

Figure 6 presents all the estimates that we obtain and also plots the distribution of κ_j and ξ_j. The autoregressive coefficient of the stochastic component is $\rho_\xi = 0.415$. The process is hardly

Figure 6: Bank returns heterogeneity



(a) Permanent, κ_j (b) Stochastic, ξ_j

Variable	Observations	Mean	SD	Min	p(25)	Median	p(75)	Max
ρ	1,378,489	0.415						
$\kappa_j + \xi_j$	1,378,489	0	0.023	-0.146	-0.009	0.002	0.013	0.076
κ_j	1,327,755	0	0.011	-0.216	-0.006	0.001	0.007	0.131
ξ_j	1,327,755	0	0.019	-0.161	-0.007	0.0012	0.01	0.137

Notes: Results of the estimation of equation 23 based on U.S. Call Reports quarterly bank-level data.

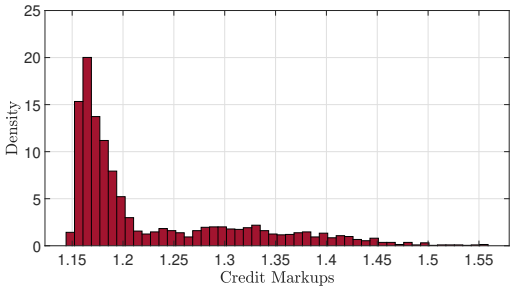
persistent and is in line with our interpretation of ξ being the "luck" component of bank profitability. Standard deviations of the permanent and stochastic components of returns heterogeneity κ_j and $\xi_{j,t}$, respectively, are 0.011 and 0.019, which stand for roughly 1 and 2 percent respectively. The bank fixed effect is very dispersed, almost to the same extent as the stochastic component. This suggests that accounting for persistent returns heterogeneity in the model could be quantitatively important.

Having estimated the parameters required to pin down the returns heterogeneity block, we employ the Tauchen and Hussey (1991) quadrature-based discretization for the stochastic part in the process 11 with σ_ξ and ρ_ξ set to our estimated values of 0.019 and 0.415, respectively. For the permanent component κ_j we assume that there is an invariant number of permanent profitability "types", whose κ_j are drown from $\mathcal{N}(0, \sigma_\kappa)$ where σ_κ is set to exactly 0.011.

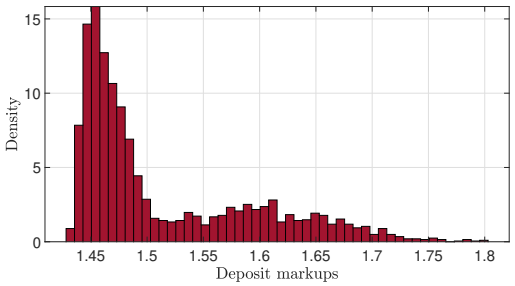
Market power There are several sets of model parameters that help determine the stationary distribution of credit and deposit market power. First, $\{\theta_k, \theta_b\}$ govern the implied homogenous credit and deposit mark-ups in the case of CES aggregation, or the mean of the respective distributions in the baseline case of Kimball systems. Second, $\{\epsilon_k, \epsilon_b\}$ help control the mark-up-size relationship and the mark-up elasticities of marginal cost shocks. We set θ_b to 10 and θ_k to 5. We set both ϵ_k and ϵ_b to 1.5. Finally, elasticity of deposit supply ν is also used for gauging the average deposit

Figure 7: Stationary distributions of bank market power

(a) Credit mark-ups



(b) Deposit mark-ups



mark-up. We set it to 1. In the data, when pooled across all quarters and banks, or by taking the average of our time-series data in Figure 1, the average credit and deposit mark-ups are 1.8 and 1.7, respectively.

Figure 7 presents the resulting cross-sectional stationary distributions of credit and deposit mark-ups. Size-weighted averages of the two objects are 1.32 and 1.6. The cross section of deposit mark-ups, which range from 1.4 to 1.8, is very much in line with the evidence in Figure 3 where values range from 1.2 to about 2. The cross section of credit mark-ups ranges from 1.15 to 1.55, which is about as dispersed as in the data but lower on average (where, as per Figure 3, mark-ups range from 1.6 to about 1.85) for the reason that we detail below.

While the model does well in pinning down the average deposit mark-up, nailing down heterogeneous credit mark-ups is more complex. It's well known in the literature that the Kimball system imposes a parameteric restriction on equilibrium relative size: $\mathrm{k}_{\max} = \theta_{\mathrm{k}}^{\frac{\theta_{\mathrm{k}}}{\epsilon_{\mathrm{k}}}}$ (Edmond et al., 2018). While lowering θ_{k} increases the steady-state average credit mark-up, it also makes the Kimball relative size constraint more likely to bind, which in turn puts a cap on bank net worth growth. This is less desirable because we want to preserve a realistic equilibrium concentration of size. For this reason we ensure that in our calibration the relative size constraint is always slack, but this comes at the cost of having a slightly lower average credit mark-up. This situation does not arise in the case of deposit mark-ups, because a sufficiently low liquidity preference parameter ν helps generating a sufficiently high aggregate deposit mark-up.[17]

Other parameters Remaining model parameters, which are listed in Table 1, are mostly standard and assigned exogenously. Periodicity in the model is quarterly. We start by discussing standard macro parameters. We set the discount factor to $\beta = 0.996$ to target a steady-state risk-free rate of 1.61%. The capital share α is set to 0.36 and the intertemporal elasticity of substitution is set to

[17] An alternative calibration approach is described in Baqaee et al. (2021) who adopt a semi-parametric strategy. In our model this approach is not computationally feasible given the incomplete markets set-up.

Table 1: Model parameterization

Parameter	Value	Description
	Macro	
β	0.996	Discounting
α	0.36	Capital Share
ψ	1	Risk Aversion
	Banking	
σ	0.9	Dividend Payout Rule
ν	1	Deposit Liquidity
λ	0.1	Leverage Constraint
ζ_1	0.01	Credit Adjustment Linear
ζ_2	1.25	Credit Adjustment Quadratic
	Bank Returns	
ρ_ξ	0.415	Idiosync. Return, Persistence
σ_ξ	0.019	Idiosync. Return, st. dev.
σ_κ	0.011	Persistent Return, st. dev.
	Bank Market Power	
θ_k	5	Demand Elasticity, Credit
θ_b	10	Supply Elasticity, Deposits
ϵ_k	1.5	Super elasticity, Credit
ϵ_b	1.5	Super elasticity, Deposits
	New Keynesian Block	
γ	10	Elasticity of Substitution, Retail
φ	100	Price Adjustment Cost, Retail
ϕ_π	1.25	Taylor Rule Inflation Coefficient
$\bar{\mathrm{R}}$	1.61	Taylor Rule Target Rate (p.a.)

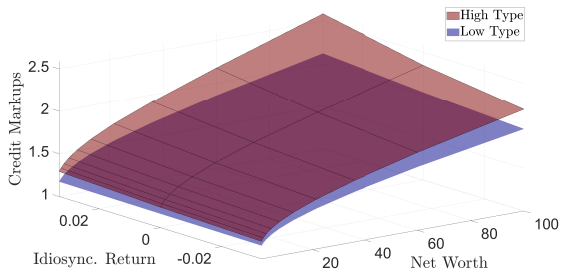
unity. Both are standard choices.

We proceed with the banking block. The dividend payout rule σ is set to 0.9, in line with Gertler et al. (2020). The leverage constraint parameter λ is 0.1, which is in the region of values used in Gertler and Kiyotaki (2010) or Gertler and Karadi (2011). We set the deposit liquidity parameter ν to a parsimonious value of 1. Credit adjustment cost parameters ζ_1 and ζ_2 are set to 0.01 and 1.25, respectively, following Jamilov (2020). These values help target the aggregate book leverage of roughly 8, in line with the empirical evidence on commercial banks (Nuno and Thomas, 2016). Parameters that govern bank returns heterogeneity $\{\rho_\xi, \sigma_\xi, \sigma_\kappa\}$ are estimated, as discussed earlier in this Section, directly from U.S. bank-level data. Similarly, the bank market power block $\{\theta_\mathrm{k}, \theta_\mathrm{b}, \epsilon_\mathrm{k}, \epsilon_\mathrm{b}\}$ is parameterized as per the discussion above.

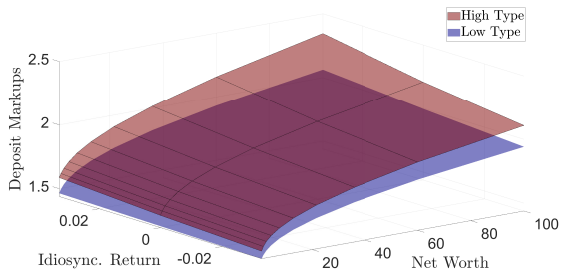
We finalize the Section with the New Keynesian block. The elasticity of substitution in the retail sector γ and the price adjustment cost are set to 10 and 100, respectively. These values are in line with the literature (Kaplan et al., 2018) and deliver a Phillips curve slope of 0.1, which is

Figure 8: Bank size, market power, and profitability

(a) Credit mark-ups



(b) Deposit mark-ups



Notes: High and low types correspond to the most and least permanently profitable (κ_j) banks in the distribution.

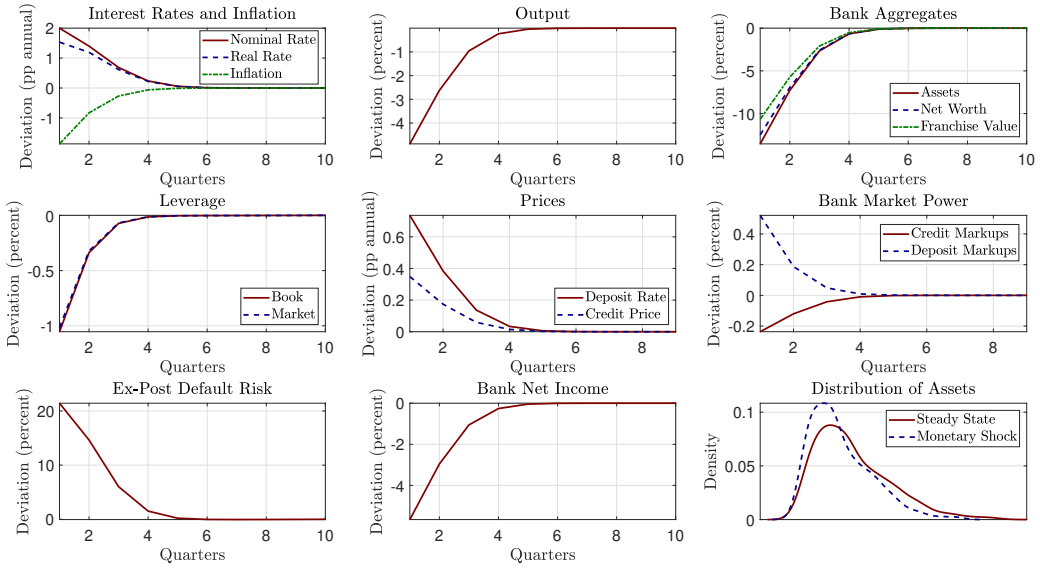
in the ballpark and slightly on the higher end of the recent micro empirical estimates Hazell et al. (2021), and an average retail mark-up of 11%. The Taylor rule inflation coefficient ϕ_π is set to 1.25 which is commonly used in New Keynesian models (Gali, 2008).

Micro banking behavior An important validation test of our framework is whether it can match the positive trilateral relationship between bank size, market power and profitability. As Figure Figure 3 in our empirical section has demonstrated, in the data it appears that large intermediaries are more profitable and charge both higher credit and deposit mark-ups. In other words, the sector features a kind of triple concentration of size, profitability, and market power.

Figure 8 plots policy surfaces for credit and deposit mark-ups $\mu^k(n, \xi, \kappa)$ and $\mu^b(n, \xi, \kappa)$ as a function of the three idiosyncratic state variables. For permanent profitability κ_j we overlay two distinct surfaces, one for the most and one for the least profitable types in the distribution. As can be seen from the Figure, both credit and deposit mark-ups are increasing in net worth as well as in both ξ and κ. In our model, the right tail of the distribution of market power is thus concentrated in the balance sheets of large and profitable intermediaries, as in the data.

This trilateral concentration that we observe in the stationary steady state is also related to the role that heterogeneity plays for the *transition dynamics*, which we analyze quantitatively in the next section. Returns heterogeneity, particularly its permanent component, creates a mass of intermediaries which are consistently profitable and large in the stationary steady state, *and* whose balance sheets are more elastic with respect to transitory monetary policy shocks.

Figure 9: Aggregate responses to a monetary policy contraction



Notes: Impulse response of aggregate variables to a 50 bps surprise positive innovation in the policy rule. Book leverage is defined as total assets over total net worth. Market leverage is defined as aggregate credit price times total assets over total net worth. Net income is defined in asset-weighted average terms. Credit mark-ups are asset-weighted averages. Deposit mark-ups are deposit-weighted averages. Franchise values, default risk, deposit rates, credit prices, and marginal costs are all asset-weighted averages. Weights (asset or deposit shares) are from the stationary steady state.

5 Monetary policy transmission

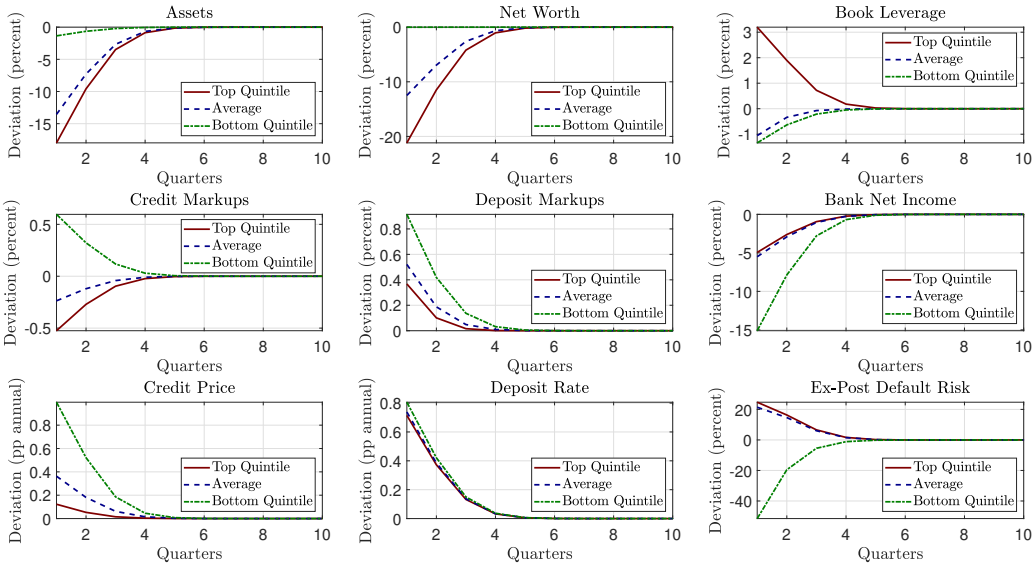
In this Section we discuss our main quantitative exercise: the response of the economy with heterogeneous banks to a temporary (one-time) unexpected monetary policy shock.

5.1 Aggregate and cross-sectional effects

We assume that before the shock hits, the economy is in the stationary steady state. We consider a quarterly innovation in the nominal interest rate of $\epsilon_{m0} = 0.5$, or 50 basis points quarterly (and 2 per cent annually). We assume that the shock reverts to the mean at the rate of 0.7. Note that this shock represents an exogenous increase in the funding costs for banks, i.e. an increase in their nominal marginal cost.

Figure 9 plots impulse responses of aggregate macroeconomic and financial variables. We see that a monetary tightening leads to a significant contraction in output and inflation. Bank assets, net worth, net income, and franchise values all fall. Crucially, monetary policy activates a significant market power channel. Our environment features real rigidities in the adjustment of credit and

Figure 10: Cross-sectional responses to a monetary policy contraction



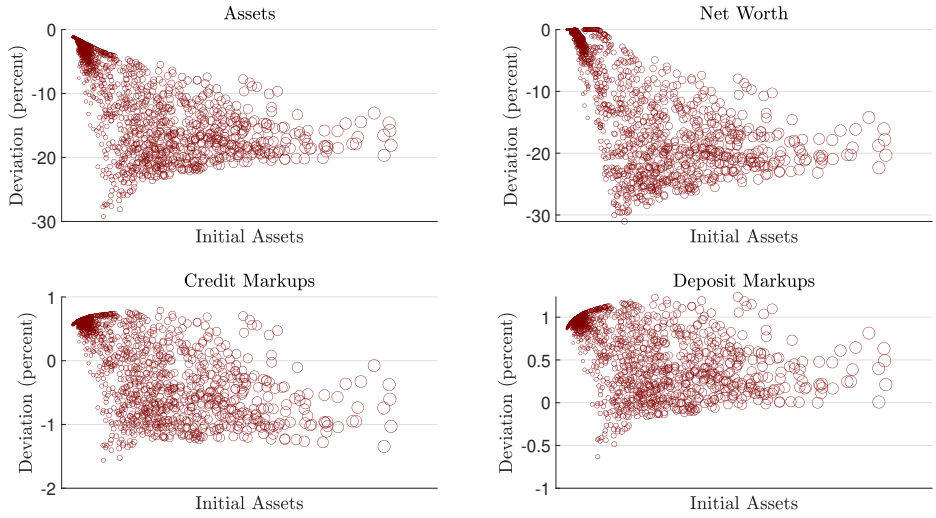
Notes: Impulse response by quintiles (20%) of the steady-state bank net worth distribution to a 50 bps surprise positive innovation in the policy rule.

deposit prices, whereby the pass-through of marginal cost shocks to market prices is smaller than under a constant-elasticity environment. Banks, on average, increase their deposit mark-ups and decrease their credit mark-ups. This is consistent with credit prices and deposit interest rates increasing less than one-to-one with the policy rate. We also observe that while leverage in the economy falls, ex-post default risk increases considerably. The final subplot of the figure depicts a considerable leftward shift in the distribution of bank assets.

The aggregate responses displayed above, however, mask a significant degree of underlying *heterogeneity*. Figure 10 displays responses across different pre-shock *quantiles* of the bank net worth distribution. The response of quantities is significantly greater for banks in the top size quintile, an observation which is consistent with our empirical findings and marginal propensity to lend increasing in initial net worth. Interestingly, we see that book leverage and ex-post default risk both become concentrated in the right tail of the size distribution, leading to an increased concentration of risk (Coimbra et al., 2022).

There is important heterogeneity in how prices and mark-ups respond across the size distribution of banks, both on the asset and the liability side. On the asset side of bank balance sheets, we observe that large (small) banks lower (increase) their credit mark-ups. Hence, credit prices respond much less (and quantities much more) for large banks, consistent with the premise that price pass-through is declining in balance sheet size. On the other hand, on the liability side of bank

Figure 11: Full distribution of bank-level responses



Notes: Impact responses to a 50bps monetary policy contraction by every intermediary in the model economy. Circle sizes are proportional to assets in the steady state.

balance sheets it is the *small* banks who increase their deposit mark-ups, and thus deposit rates, by more. Heterogeneous, two-sided price rigidity is a key feature of our environment, pointing to the importance of the interaction between bank market power and balance sheet size as a monetary policy transmission channel. Note that the relatively higher (lower) response of credit (deposit) mark-ups by large banks is also in line with our empirical findings.

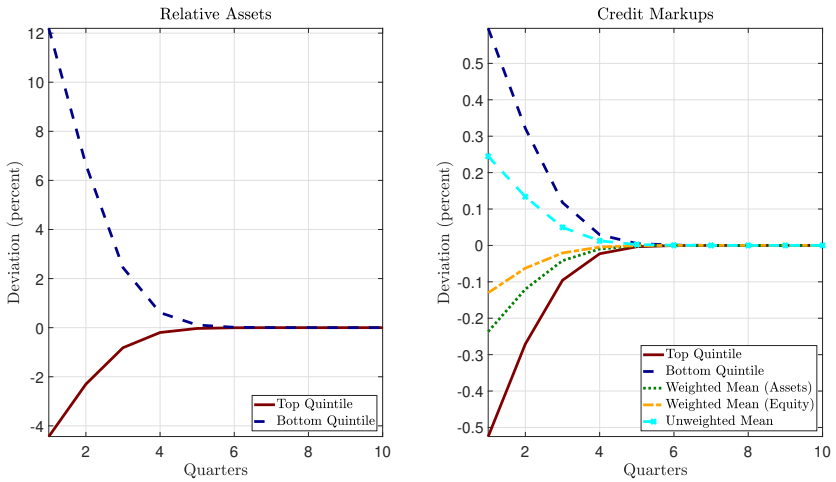
In order to further illustrate the full distributional effects of monetary policy, Figure 11 shows how every single intermediary in our economy responds (on impact) to the policy shock. In terms of balance sheet quantities, we see that every bank in the economy contracts its size. For credit mark-ups, the distribution compresses from both tails - ex-ante small banks raise mark-ups while ex-ante large banks lower them, thus leading to a decline in dispersion. For deposit mark-ups, the distribution of responses is almost universally positive and uniform. And because in the cross section deposit mark-ups are increasing with size, dispersion increases following the shock.[18]

5.2 Inspecting the mechanism

In this section we conduct an in-depth analysis of the aggregate and heterogeneous responses presented above. We try to further shed light on the heterogeneity in price rigidities, the role of returns heterogeneity, and the role that credit and deposit market power play individually. We

[18]In other words, small banks display a larger *percentage* point increase in deposit mark-ups than the large ones. However, the opposite holds for the *absolute* change.

Figure 12: Relative size and credit mark-ups



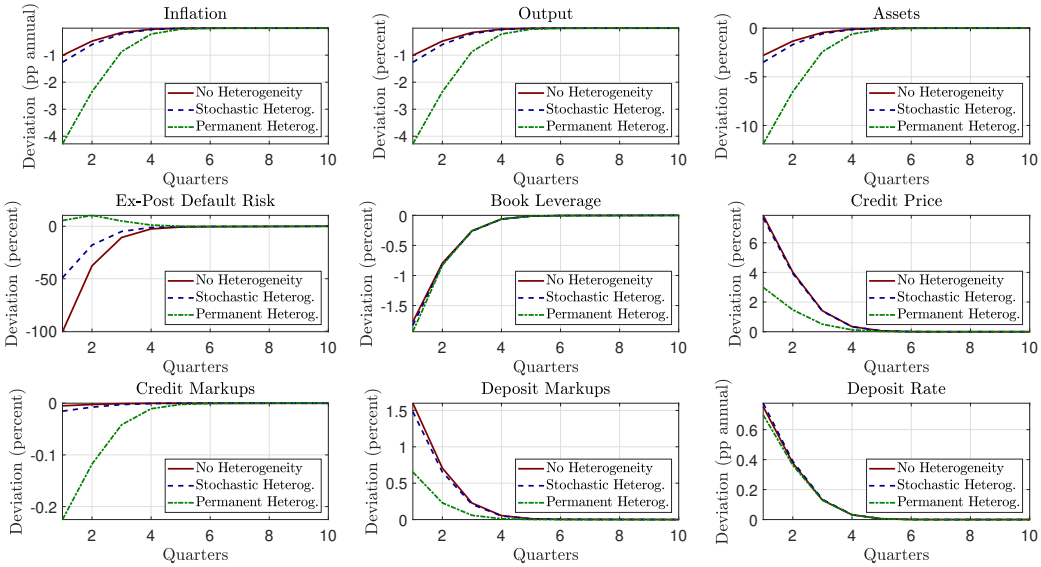
Notes: Impact responses of relative asset and credit mark-ups to a 50bps monetary policy contraction.

also decompose the total response of output into various partial and general equilibrium channels. Finally, we check whether the real rigidity effect is asymmetric by quantifying responses to an expansionary monetary policy shock.

Understanding the heterogeneous response of credit mark-ups. We highlight an important observation from Figure 10 that large (small) banks lower (raise) their credit mark-ups following the monetary contraction. This potentially puzzling result can be explained by the fact that it is banks' *relative size* which matters for the differential behavior of mark-ups in the cross-section. Figure 12 shows the transitional behavior of relative assets and mark-ups. In response to the monetary contraction, banks in the top asset quintile become relatively *smaller* while the ones in the bottom quintile relatively *larger*. In other words, dispersion as well as concentration fall. Since the dynamic of relative size is a sufficient statistic for the credit mark-up response, the differential response of mark-ups is no more surprising. This same figure also shows that aggregation and weighting matters for the case of credit mark-ups. The average credit mark-up *rises* if it is measured in terms of the unweighted mean, whereas the average mark-up falls if the mean is asset- or equity-weighted. For consistency with every other variable in the model and in our empirical analysis, we settle on the weighted instead of the unweighted mean, however highlighting this nuance is important. Weighting does not matter for the case of deposit mark-ups.

The role of bank heterogeneity Next we investigate quantitatively the importance of bank returns heterogeneity for monetary policy transmission. We compare the response of the same variables in *three* distinct economies: (i) no permanent κ_j or stochastic ξ_j returns heterogeneity, (ii) only

Figure 13: The role of bank heterogeneity



Notes: Impact responses to a 50bps monetary policy contraction: no bank heterogeneity (solid), only idiosyncratic heterogeneity (dashed), only permanent heterogeneity (dotted).

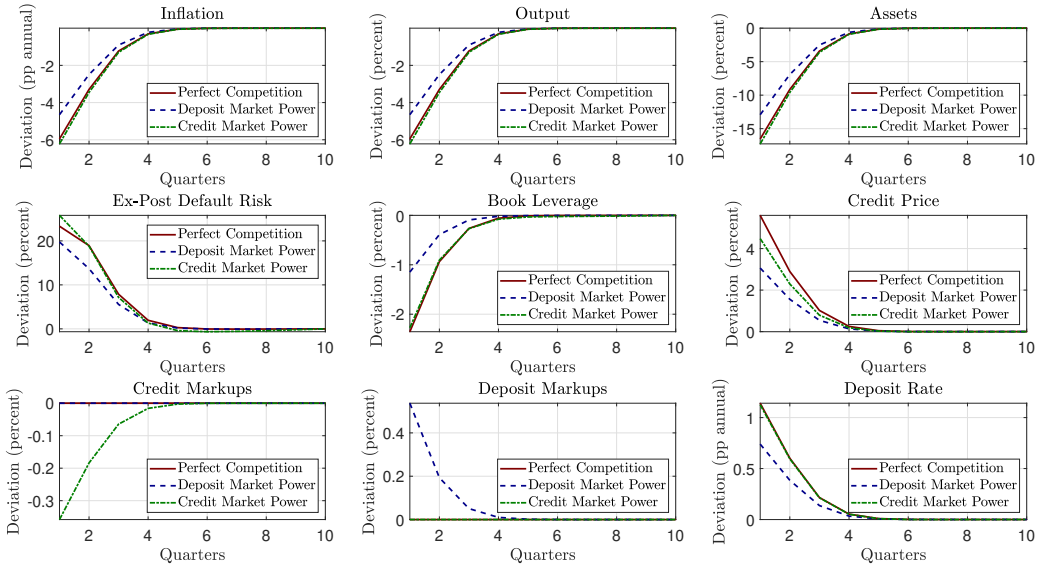
stochastic heterogeneity, and (iii) only permanent heterogeneity.[19]

Figure 13 plots the results. We see immediately that permanent returns heterogeneity plays a key role in shaping the transmission of monetary policy for aggregate output, inflation, and the financial sector. Permanent heterogeneity leads to significant amplification of monetary policy shocks. This is due to the combination of two effects. First, larger banks have a substantially higher MPL. Second, larger banks have a higher credit and a lower deposit mark-up elasticity. A higher aggregate MPL of the economy with permanent heterogeneity leads to greater first-degree amplification of bank balance sheets. On top of that, initializing from the steady state with a higher fraction of very profitable and large banks means that the aggregate response of credit (deposit) mark-ups will be much greater (smaller). And since credit and deposit market price rigidities act as an amplifier (dampener) of monetary shocks, this second-level market power channel leads to yet further amplification.

The role of two-sided bank market power. In order to identify the roles played by credit and deposit market power *individually*, we compare the response of economic and financial aggregates in three economies: perfect competition, deposit market power only, and credit market power only.

[19]In the "idiosyncratic only" economy we set σ_κ to 0. In the "permanent only" economy we lower σ_ξ by 25%, which ensures that average return on assets and net worth growth are both positive. In the "no heterogeneity" economy we set σ_κ to 0 and lower σ_ξ by 25%.

Figure 14: The role of two-sided bank market power



Notes: Impact responses to a 50bps monetary policy contraction: perfectly competitive markets (solid), deposit market power only (dashed), credit market power only (dotted).

In the "perfect competition" economy we turn off the deposit liquidity (by setting ν to infinity) and the real rigidity (by setting θ_k and θ_b to infinity) channels. In the "deposit market power" economy we set $\theta_k = \infty$. Finally, in the "credit market power" economy we set $\theta_b = \infty$ and $\nu = \infty$.

Figure 14 presents the results of this experiment. Credit market power amplifies (albeit mildly) while deposit market power dampens the response to the same monetary policy shock on output, inflation and balance sheet variables. The amplifying effect of credit market power stems from imperfect credit market competition and non-CES demand. Because credit mark-ups are endogenous and pro-cyclical, prices are rigid and respond less than one-to-one to marginal cost movements. Imperfect pass-through is particularly present in our environment with heterogeneity, since the credit demand elasticity σ_j (pass-through) is increasing (decreasing) with relative bank size. As a result, quantities are more elastic with respect to monetary surprises in the short run. Note also that in our set-up capital quantites and prices move in *opposite* directions, akin to a leftward shift in the aggregate credit supply curve that moves in response to an increase in the banks' aggregate cost of funds.

On the other hand, non-CES deposit supply reduces the pass-through of the risk-free rate shock onto the deposit rate and thus the marginal cost itself. This occurs for two reasons. First, the Kimball deposit mark-up elasticity is increasing in the risk-free rate. Second, liquidity preferences

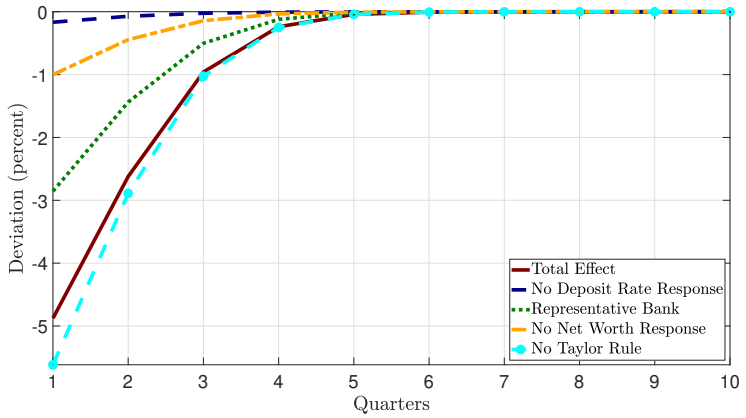
intensify following monetary contractions as the $\frac{C_t^{\psi}}{B_t^{\nu}}$ term in Equation 5 is countercyclical. A higher real rate makes deposits more attractive for households as the marginal utility of deposit holdings grows, and banks internalize this effect by exercising greater market power endogenously. As a result, the deposit spread increases, the pass-through onto marginal costs is incomplete, and the response of balance sheet quantities is of smaller magnitude.

Another noteworthy observation that stems from Figure 14 is related to the trade-off between financial stability and competition. Note that in all three economies, ex-post default risk increases following the monetary contraction. However, in the economy with deposit market power only, the response is smaller by roughly 5 percentage points on impact and considerably more cumulatively. Equilibrium deposit mark-ups are clearly welfare-reducing because households face deposit rates that are lower than in the first-best perfect competition case. Additionally, deposit mark-ups are counter-cyclical. On the other hand, monetary policy contractions lead to a lesser deterioration in financial stability in that same deposit market power economy. In other words, the monetary authority potentially faces a trade-off between bank default risk on one side and deposit market power on the other. The competition-stability view has a long-standing tradition in the literature (Keeley, 1990; Hellman et al., 2000; Repullo, 2004; Beck et al., 2006; Carlson et al., 2022). Although we abstract from quantitative normative analysis, this is an important highlight for future research.

Decomposition of the output response In our economy, the total response of aggregate macro variables to a monetary policy shock is comprised of several layers of partial and general equilibrium channels. We now quantify these channels by decomposing the total response of output in the baseline economy into four components. First, we do not allow deposit interest rates to respond to the shock. In the absence of this mechanism, monetary policy affects the financial sector only through the household's marginal rate of substitution, which acts as the banks' (augmented) pricing kernel. Second, we compute the response of a "representative bank" whose profitability is equal to average R^T in the distribution. This should help address the potential concern that our heterogeneity is uninformative, i.e., that the response of the average intermediary in our setup could approximate the baseline economy well. Third, we don't allow bank net worth to respond at all, thus shutting down the whole financial accelerator mechanism. Finally, we don't allow for any feedback effects between the economy and the Taylor rule via the central bank's endogenous reactions to inflation caused by the initial monetary shock.

Figure 15 presents the result of this exercise. First, we see that the banks' cost of funds, and not the households' pricing kernel, is essential for the bank lending channel of monetary transmission to operate. Shutting down the deposit rate channel eliminates more than 95% of the output response. Second, output response in the counterfactual economy with a representative bank is roughly 40%

Figure 15: Decomposition of the output response



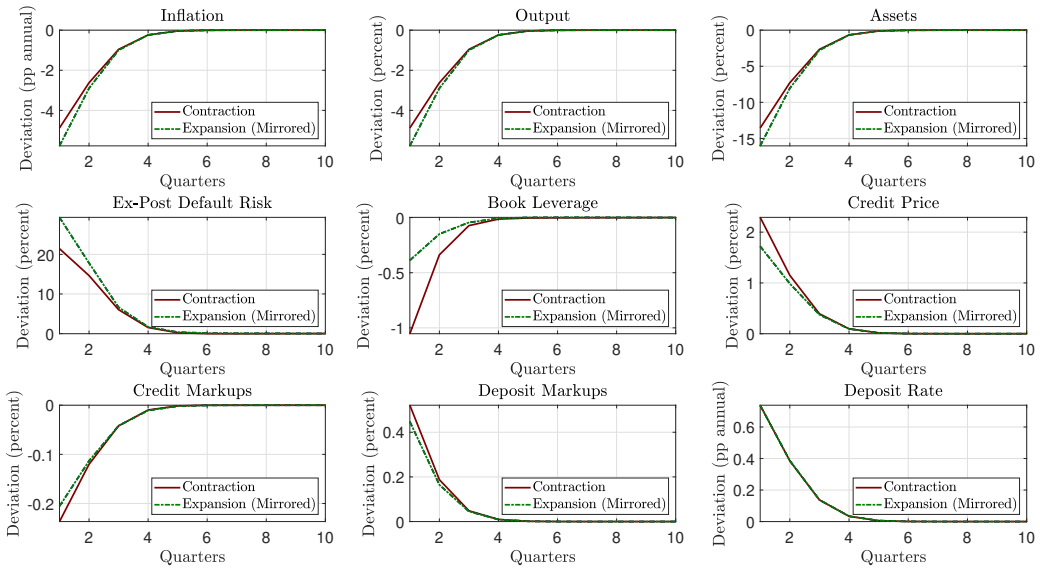
Notes: Impulse response of aggregate output under alternative counterfactual scenarios.

smaller than in the baseline. This validates our premise that returns heterogeneity cannot be approximated well by profitability of the "average" bank. Properly accounting for returns and size heterogeneity is essential for capturing and quantifying the full extent of the bank lending channel. Third, shutting down the whole financial accelerator mechanism reduces the output response by 80%. This is not surprising, considering that our model is essentially one with a single factor of production. If bank net worth is not allowed to adjust, much of the action in terms of the response of either quantities or prices is turned off because all those choices are functions of the only "endogenous" idiosyncratic state, i.e., net worth. Finally, shutting down the Taylor rule inflation feedback effect amplifies the response of output only mildly.

Asymmetric responses In our framework monetary policy can have sizeable non-linear (asymmetric) effects. Given our non-linear solution method, we should be able to capture asymmetries that are generated by the Kimball "quasi-kinks" in both credit demand and deposit supply curves. To this extent, we simulate a 50 basis point monetary *expansion* in the otherwise unchanged baseline economy.

Figure 16 plots the result where all responses to the expansionary shock are mirrored (flipped). Notice how equilibrium asymmetries are very strong for both output and inflation. In our economy, monetary policy expansions are 18 percent more impactful on output and 13 percent more impactful on inflation. This is due to the fact that prices (particularly in the credit market) are significantly more rigid when adjusting downwards rather than upwards. Quantities, on the other hand, are more elastic following monetary expansions. In other words, the real rigidity effect is asymmetric; it is particularly pronounced in the case of a *fall* in marginal costs. And as Linde and Trabandt (2018) show, this asymmetry effect shows up only when solving the model non-linearly, as it is the case

Figure 16: Monetary contraction vs expansion



Notes: Impulse responses to a 50bps monetary policy shock: contraction (solid) vs expansion (dashed).

in our paper. This is due to (the change in) the optimal relative price being convex in (the change in) the marginal cost. Hence the pass-through from prices to marginal costs is larger if the percent change in marginal cost is positive (i.e., a rise in the monetary policy rate) rather than negative (a fall in the policy rate).

5.3 Higher moments, two-sided CES, and expansionary shocks in the cross-section

In Appendix B.3 we supplement our main findings with several additional results. First, we compute the responses of second (standard deviation) and third (skewness and Herfindahl index) moments of key financial variables to a contractionary monetary policy shock. Second, we compute the transitional dynamics in economies with two-sided CES demand instead of Kimball demand. Finally, we show heterogeneous responses to an expansionary monetary shock.

6 Conclusion

We have developed a New Keynesian model with heterogeneous banks, incomplete insurance, two-sided market power, and nominal rigidities (aka HBANK). The model incorporates advances

from the literature on heterogeneous agents on one side and imperfect competition on the other. Our analysis is motivated and validated by detailed micro-level evidence from the U.S. banking sector. We have shown that endogenous *market power* on both the asset and the liability side, as well as *heterogeneity* in the banking cross-section, are crucial features to account for the transmission of monetary policy to the real economy through the credit system. Future work on the monetary policy transmission mechanism via the banking sector should develop at least along three main directions. First, a deeper understanding of how bank heterogeneity and imperfect competition interact in a Zero-Lower-Bound (ZLB) interest rate environment. Second, an analysis -within a HBANK framework- of the effects of unconventional monetary policies, such as forward guidance, asset purchase programs, and negative interest rates. Third, introducing empirically-consistent equilibrium house price movements as drivers of bank market power and risk-taking.

References

Adrian, T. and N. Boyarchenko (2015): "Intermediary leverage cycles and financial stability," *Staff Reports, Federal Reserve Bank of New York*, 567.

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

Altavilla, C., M. Boucinha, and J.-L. Peydró (2018): "Monetary policy and bank profitability in a low interest rate environment," *Economic Policy*, 33, 531–586.

Amiti, M. and D. Weinstein (2018): "How Much Do Idiosyncratic Bank Shocks Affect Investment? Evidence from Matched Bank-Firm Data," *Journal of Political Economy*, 126.

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

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

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

Begenau, J., S. Bigio, J. Majerovitz, and M. Vieyra (2021): "A Q-Theory of Banks," *Manuscript*, NBER Wokring Paper 27935.

Begenau, J. and T. Landvoigt (2022): "Financial Regulation in a Quantitative Model of the Modern Banking System," *Review of Economic Studies*, Forthcoming.

Benetton, M. (2021): "Leverage Regulation and Market Structure: A Structural Model of the U.K. Mortgage Market," *The Journal of Finance*, 76, 2997–3053.

Benhabib, B., A. Bisin, and M. Luo (2019): "Wealth distribution and social mobility in the US: A quantitative approach," *American Economic Review*, 109(5).

Bianchi, J. and S. Bigio (2022): "Banks, Liquidity Management and Monetary Policy," *Econometrica*, 90(1).

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

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

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

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

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

Carlson, M., S. Correia, and S. Luck (2022): "The Effects of Banking Competition on Growth and Financial Stability: Evidence from the National Banking Era," *Journal of Political Economy*, 130.

Coimbra, N., D. Kim, and H. Rey (2022): "Central Bank Policy and the concentration of risk: Empirical estimates," *Journal of Monetary Economics*.

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

Corbae, D. and P. D'Erasmo (2020): "Rising bank concentration," *Journal of Economic Dynamics and Control*, 115, 103877.

——— (2021): "Capital Buffers in a Quantitative Model of Banking Industry Dynamics," *Econometrica*, 89, 2975–3023.

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

De Loecker, J. and J. Eeckhout (2021): "Global Market Power," *Working Paper*.

De Loecker, J., J. Eeckhout, and G. Unger (2020): "The Rise of Market Power and the Macroeconomic Implications," *The Quarterly Journal of Economics*, 135, 561–644.

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

——— (2021): "Banking on Deposits: Maturity Transformation without Interest Rate Risk," *Journal of Finance*, 76.

Edmond, C., V. Midgrigan, and D. Xu (2018): "How Costly Are Markups?" *NBER Working Paper 24800*.

Egan, M., A. Hortacsu, and G. Matvos (2017): "Deposit Competition and Financial Fragility:

Evidence from the US Banking Sector," *American Economic Review*, 107(1).

Farhi, E. and J. Tirole (2021): "Shadow Banking and the Four Pillars of Traditional Financial Intermediation," *The Review of Economic Studies*, 88(6).

Farhi, E. and I. Werning (2016): "A Theory of Macroprudential Policies in the Presence of Nominal Rigidities," *Econometrica*, 84(5).

Galaasen, S., R. Jamilov, R. Juelsrud, and H. Rey (2021): "Granular Credit Risk," *NBER Working Paper 27994*.

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

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

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

——— (2015): "Monetary Policy Surprises, Credit Costs, and Economic Activity," *American Economic Journal: Macroeconomics*, 7, 44–76.

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

——— (2015): "Banking, Liquidity, and Bank Runs in an Infinite Horizon Economy," *American Economic Review*, 105(7), 2011–2043.

Gertler, M., N. Kiyotaki, and A. Prestipino (2016): "Wholesale Banking and Bank Runs in Macroeconomic Modelling of Financial Crises," *Handbook of Macroeconomics*, 2.

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

Gurkaynak, R., B. Sack, and E. Swanson (2005): "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements," *International Journal of Central Banking*, 1(1).

Haddad, V. and T. Muir (2021): "Do Intermediaries Matter for Aggregate Asset Prices?" *Journal of Finance*, 76.

Hazell, J., J. Herreno, E. Nakamura, and J. Steinsson (2021): "The Slope of the Phillips Curve: Evidence from U.S. States," *Quarterly Journal of Economics*, Forthcoming.

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

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

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

Jamilov, R. and T. Monacelli (2020): "Bewley Banks," *CEPR Discussion Paper No. DP15428*, Available at SSRN: https://ssrn.com/abstract=3737561.

Jarociński, M. and P. Karadi (2020): "Deconstructing Monetary Policy Surprises - The Role of Information Shocks," *American Economic Journal: Macroeconomics*, 12, 1–43.

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

Kaplan, G., K. Mitman, and G. Violante (2020): "The Housing Boom and Bust: Model Meets Evidence," *Journal of Political Economy*, 128.

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

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

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

Klenow, P. and J. Willis (2016): "Real Rigidities and Nominal Price Changes," *Economica*.

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

Linde, J. and M. Trabandt (2018): "Should we use linearized models to calculate fiscal multipliers?" *Journal of Applied Econometrics*, 33.

Mertens, K. and M. O. Ravn (2013): "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States," *American Economic Review*, 103, 1212–47.

Miranda-Agrippino, S. and H. Rey (2020): "U.S. Monetary Policy and the Global Financial Cycle," *The Review of Economic Studies*, 87, 2754–2776.

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

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

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

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

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

Scharfstein, D. and A. Sunderam (2016): "Market Power in Mortgage Lending and the Transmission of Monetary Policy," *Working Paper*.

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

Stock, J. H. and M. W. Watson (2012): "Disentangling the Channels of the 2007-09 Recession," *Brookings Papers on Economic Activity*, 1, 81–135.

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

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

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

Wang, Y., T. M. Whited, Y. Wu, and K. Xiao (2020): "Bank Market Power and Monetary Policy Transmission: Evidence from a Structural Estimation," *NBER Working Paper 27258*.

Whited, T. M., Y. Wu, and K. Xiao (2021): "Low interest rates and risk incentives for banks with market power," *Journal of Monetary Economics*, 121, 155–174.

Woodford, M. (2003): "Interest and Prices: Foundations of a Theory of Monetary Policy," *Princeton University Press*.

Online Appendix for "HBANK: Monetary Policy with Heterogeneous Banks"

Marco Bellifemine Rustam Jamilov Tommaso Monacelli

March 20, 2022

Contents

A Empirical appendix **2**
A.1 Response of higher order moments to monetary shocks . . . 2
A.2 Robustness . . . 3
A.3 Response of macro variables . . . 10
A.4 Data construction . . . 11
A.5 Credit and deposit mark-up estimation . . . 12

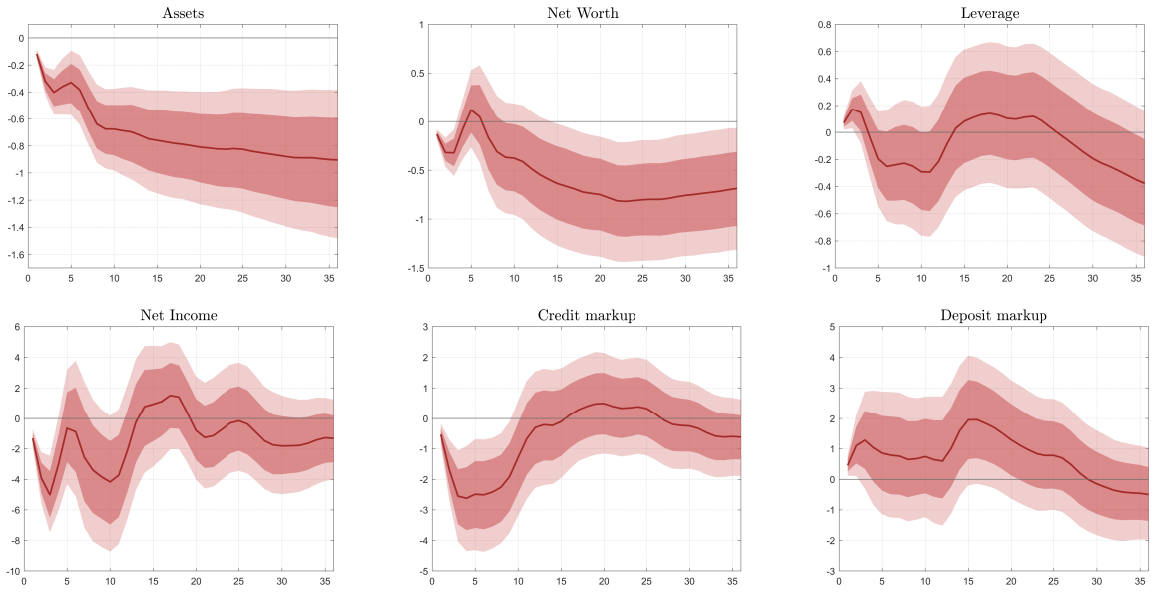
B Model appendix **17**
B.1 Derivation of the household problem . . . 17
B.2 Klenow-Willis specification . . . 18
B.3 Additional model results . . . 19
B.4 Solution algorithms . . . 22

A Empirical appendix

A.1 Response of higher order moments to monetary shocks

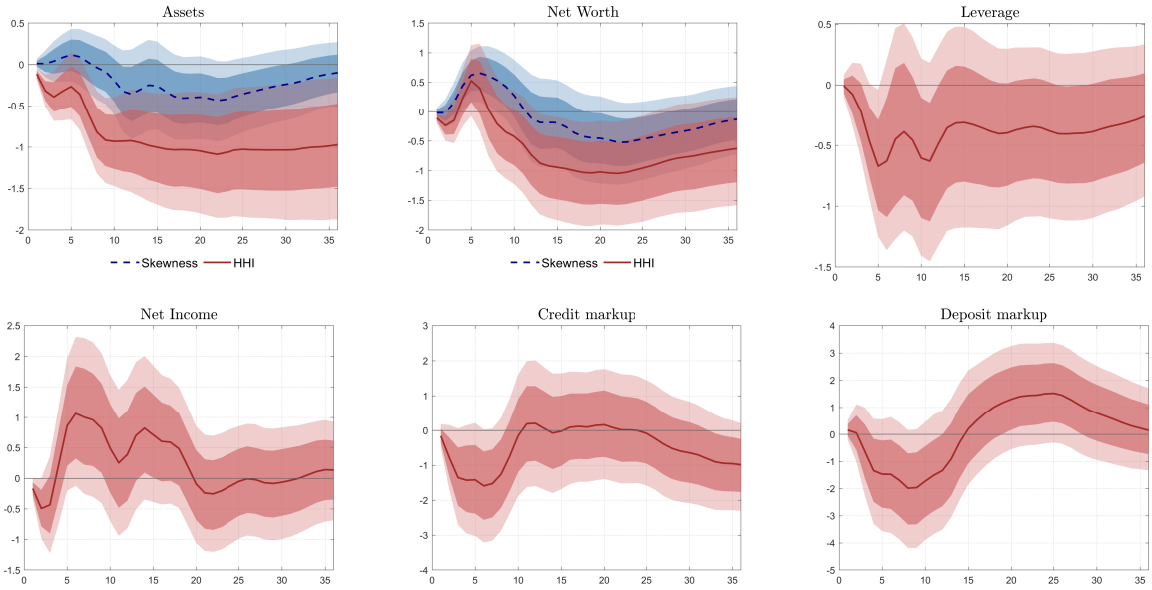
Figures A.1 and A.2 shed further light on the heterogeneous response of the banking sector to monetary shocks, by plotting the IRFs of standard deviation and skewness of our banking variables. We see that, in line with the quintile response analyzed in Section 2.2, the dispersion of both assets and net income declines in response to the monetary tightening, as big banks shrink their size by more. Accordingly, there is a sharp decrease in concentration of both assets and net worth as measured by the Herfindahl-Hirschman index, while skewness does not show any significant response. We document a decrease in both the standard deviation and –to a lesser extent– the skewness of credit mark-ups. This fits well into the picture we have drawn in Section 2. On the one hand, big banks display larger credit mark-ups. At the same time, however, they also reduce mark-ups by more, relatively to small banks, in response to monetary shocks. As a result, a monetary tightening disproportionally affects the right tail of the credit mark-up distribution, hence decreasing both the dispersion and skewness of the density. When it comes to the deposit mark-up, instead, it is more difficult to grasp the behavior of both the second and third moment of the distribution, as the estimated responses of both moments are noisy, and never significantly different from zero.

Figure A.1: Response of banking variables to a monetary shock: second moment



Notes. We use unweighted standard deviation for assets and net worth; asset-weighted standard deviation for leverage, net income, and credit mark-ups; and deposit-weighted standard deviation for deposit mark-ups. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

Figure A.2: Response of banking variables to a monetary shock: third moment



Notes. We use unweighted skewness and HHI for assets and net worth; asset-weighted skewness for leverage, net income, and credit mark-ups; and deposit-weighted skewness for deposit mark-ups. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

Table A.1 shows the F statistics for the first-stage IV. We see that all our specifications satisfy standard instrument relevance requirements, as they are all well above the rule of thumb threshold of 10 proposed in Stock et al. (2002).

Table A.1: First stage F-statistics from VAR: higher order moments

	Standard Deviation	Skewness	HHI
Assets	31.69	35.06	34.72
Net Worth	35.29	35.71	37.50
Leverage	29.06	32.83	-
Net Income	37.59	35.36	-
Credit Mark-up	36.04	35.23	-
Deposit Mark-up	36.30	37.39	-

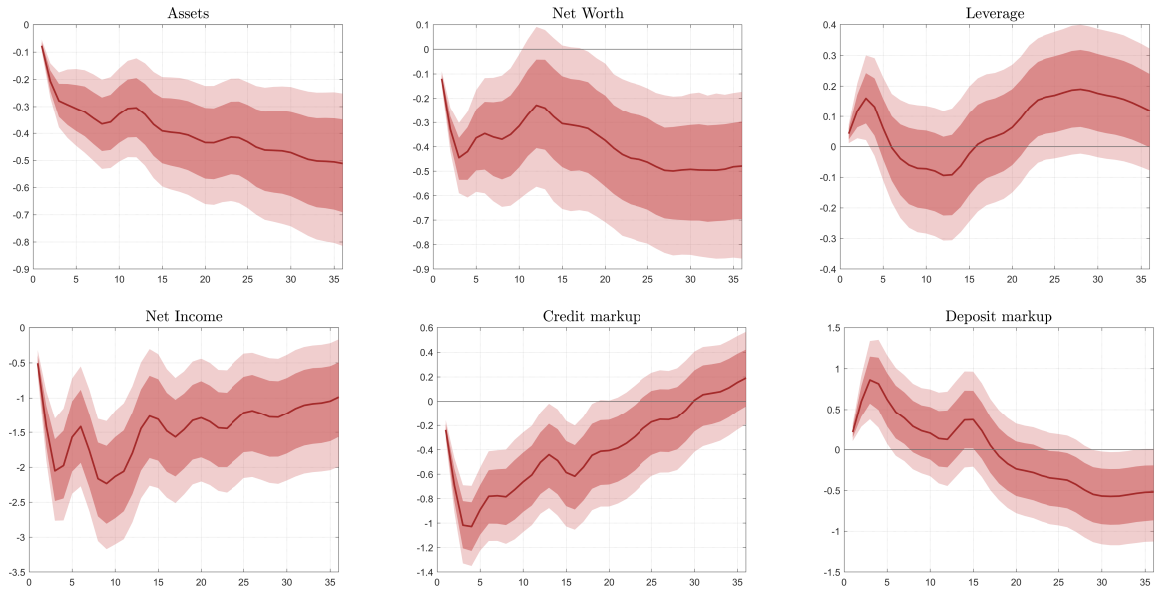
A.2 Robustness

In this section, we show that the responses to monetary shocks we estimate in Section 2.2 are robust to a variety of different specifications.

Figures A.3 and A.4 show that all our estimated aggregate and heterogeneous responses are

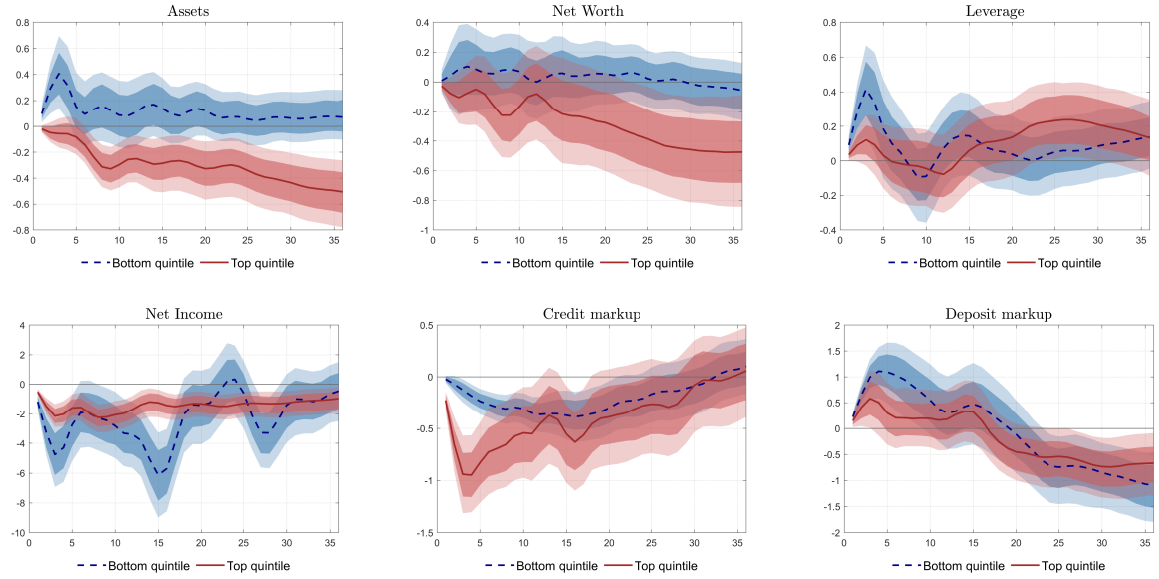
unchanged –both qualitatively and quantitatively– when we instrument monetary shocks using the series proposed in Jarociński and Karadi (2020), which accounts for the information content of FED's announcements.

Figure A.3: Controlling for the information content: aggregate responses



Notes. We use totals for assets, equity, leverage, and net income; asset weighted average for credit mark-ups; and deposit weighted average for deposit mark-ups. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

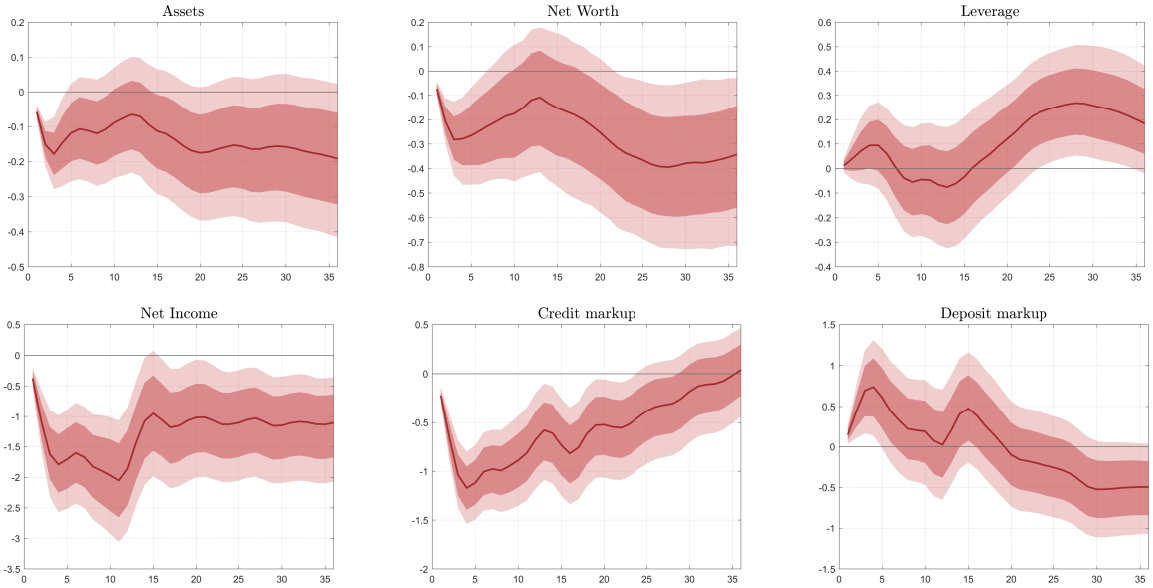
Figure A.4: Controlling for the information content: heterogeneous responses



Notes. Assets, net worth, leverage, and net income are totals within the quintile. Credit mark-ups are asset weighted within the quintile. Deposit mark-ups are deposit weighted within the quintile. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

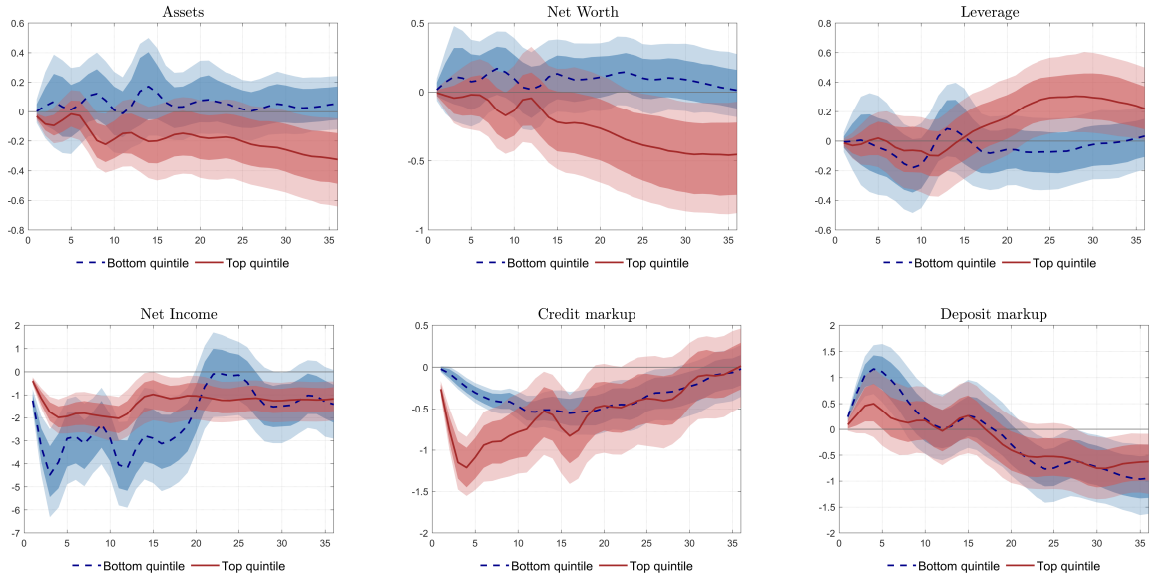
Figures A.5 and A.6 and Figures A.7 and A.8 plot estimated aggregate and heterogeneous IRFs when we estimate the reduced-form VAR coefficients over the restricted samples 1990:02-2017:12 and 1985:01-2012:06 respectively. Again, in both cases all our results are materially unaffected.

Figure A.5: Starting the sample in 1990m1: aggregate responses



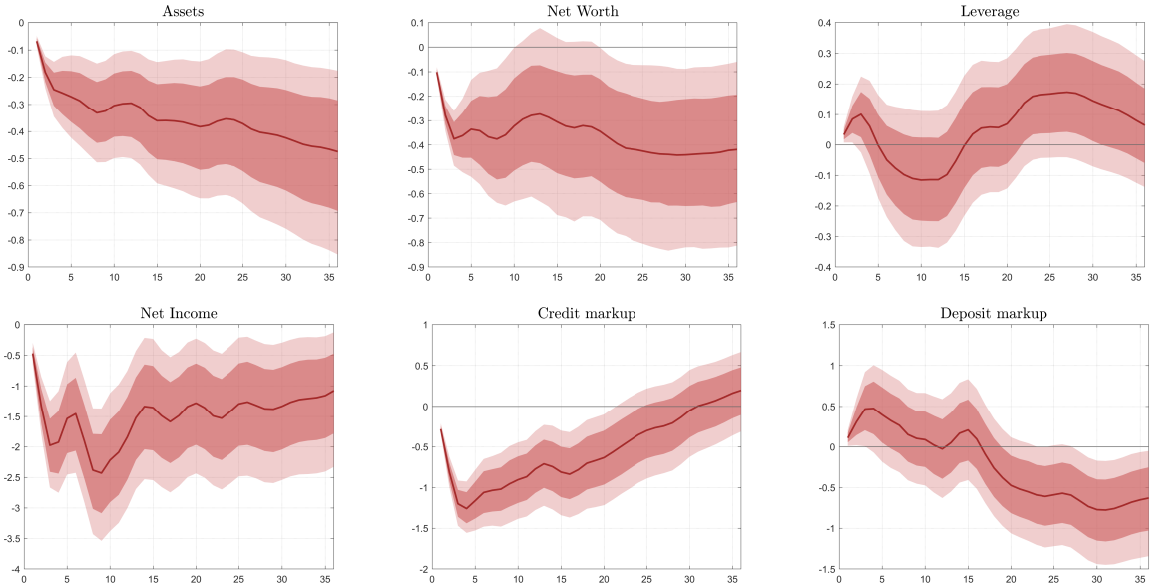
Notes. We use totals for assets, equity, leverage, and net income; asset weighted average for credit mark-ups; and deposit weighted average for deposit mark-ups. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

Figure A.6: Starting the sample in 1990m1: heterogeneous responses



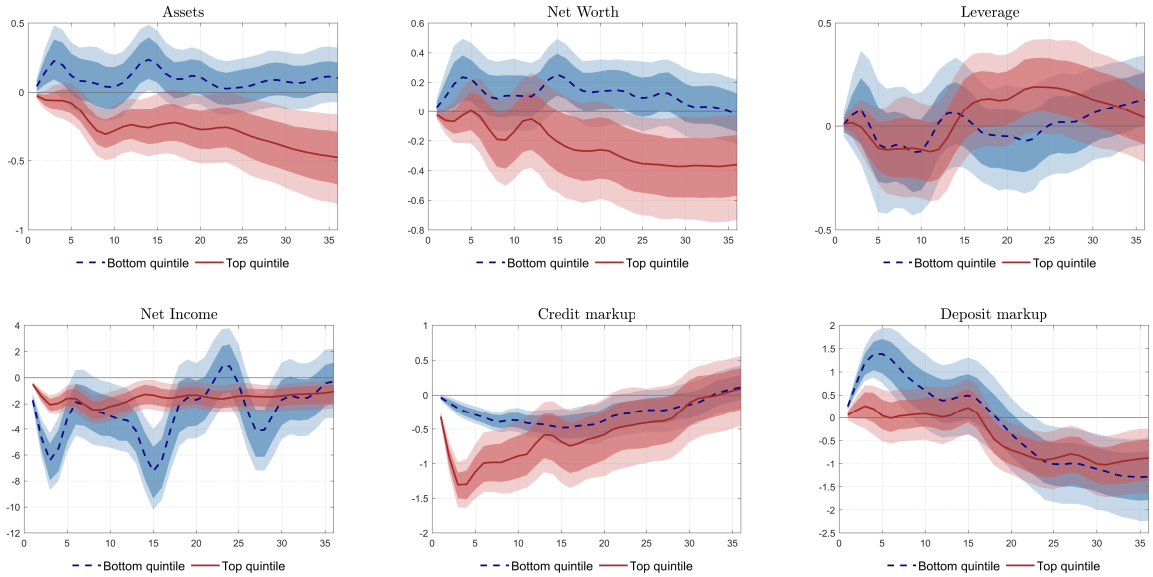
Notes. Assets, net worth, leverage, and net income are totals within the quintile. Credit mark-ups are asset weighted within the quintile. Deposit mark-ups are deposit weighted within the quintile. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

Figure A.7: Ending the sample in 2012m6: aggregate responses



Notes. We use totals for assets, equity, leverage, and net income; asset weighted average for credit mark-ups; and deposit weighted average for deposit mark-ups. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

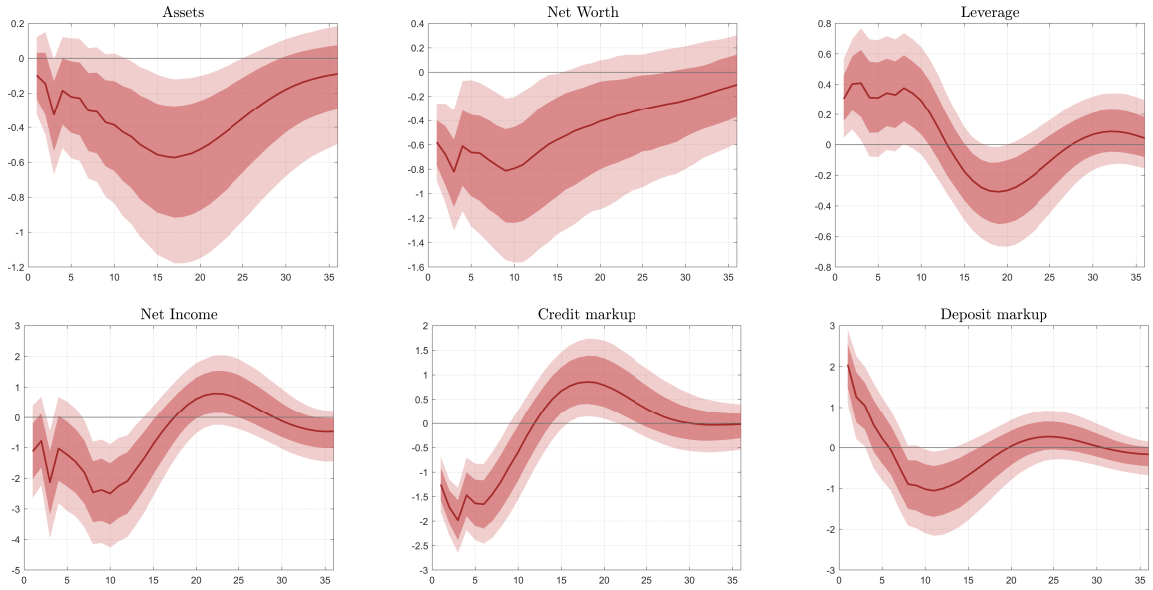
Figure A.8: Ending the sample in 2012m6: heterogeneous responses



Notes. Assets, net worth, leverage, and net income are totals within the quintile. Credit mark-ups are asset weighted within the quintile. Deposit mark-ups are deposit weighted within the quintile. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

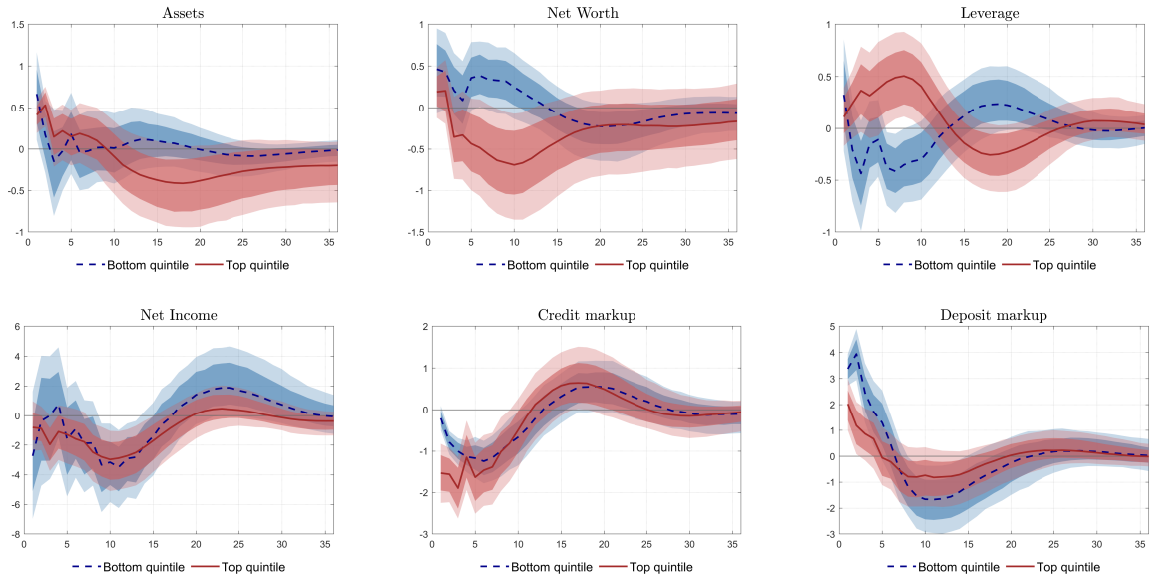
Finally, Figures A.9 and A.10 plot aggregate and heterogeneous IRFs based on quarterly data. Even though we estimate aggregate responses more imprecisely, the results are qualitatively in line with what we find in the baseline specifications. Moreover, the heterogeneous response of net worth and credit and deposit mark-ups is qualitatively in line with what we describe in the main text. Specifically, compared to small banks, banks in the top 20% of the size distribution experience a larger decrease in their net worth and credit mark-up, while they display a smaller increase in the deposit mark-up. On the other hand, when estimated at a quarterly frequency, the heterogeneous responses of assets and net income are too imprecise to show any clear pattern.

Figure A.9: Quarterly data: aggregate responses



Notes. We use totals for assets, equity, leverage, and net income; asset weighted average for credit mark-ups; and deposit weighted average for deposit mark-ups. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

Figure A.10: Quarterly data: heterogeneous responses



Notes. Assets, net worth, leverage, and net income are totals within the quintile. Credit mark-ups are asset weighted within the quintile. Deposit mark-ups are deposit weighted within the quintile. Lightly and darkly shaded areas represent respectively 90% and 68% confidence intervals.

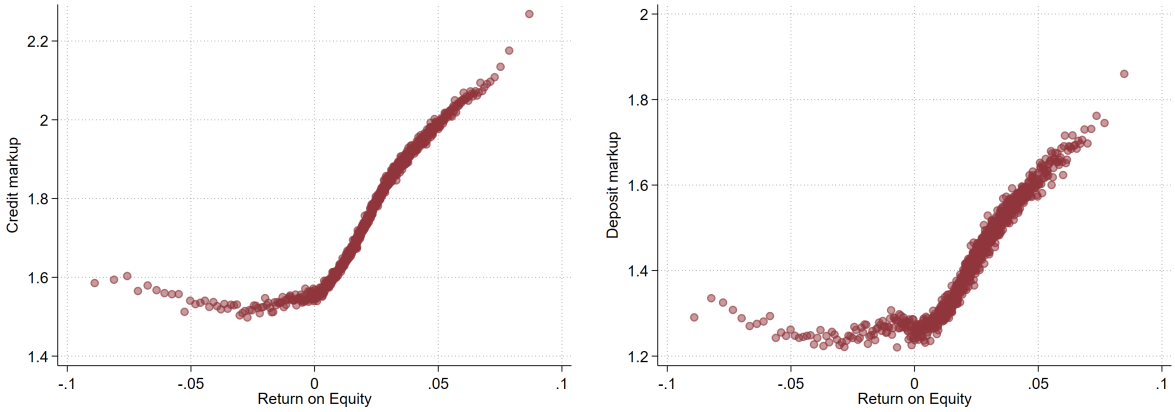
Table A.2 reports the first stage IV statistics of from regressing our reduced-form VAR residual for the fed funds rate onto the instrument for the monetary shock. None of our specifications raises

concerns of weak instruments, as all the F statistics are well above the rule of thumb threshold of 10 proposed in Stock et al. (2002).

Table A.2: First stage F-statistics from VAR

	Baseline	Karadi-Jarociński	Start 1990	End 2012	Quarterly
Aggregate responses					
Assets	36.38	18.18	34.79	30.16	35.78
Net Worth	33.10	16.08	32.10	26.84	33.62
Leverage	31.70	13.96	30.28	26.15	32.55
Net Income	38.02	19.13	31.87	30.89	40.36
Credit Mark-up	35.44	17.00	32.61	27.82	39.33
Deposit Mark-up	32.89	18.21	32.49	26.95	36.04
Heterogeneous responses					
Assets Q1	33.60	17.34	30.46	27.47	35.61
Assets Q5	35.65	17.30	34.24	29.03	35.97
Net Worth Q1	33.52	17.65	31.67	26.60	35.47
Net Worth Q5	33.54	17.34	32.96	28.06	33.80
Leverage Q1	33.41	18.61	29.50	27.51	35.90
Leverage Q5	33.43	15.79	31.05	27.61	34.00
Net Income Q1	32.21	19.51	30.34	26.17	36.45
Net Income Q5	31.25	12.73	33.32	25.61	36.41
Credit Mark-up Q1	36.10	20.77	33.13	29.66	38.63
Credit Mark-up Q5	35.27	16.81	32.58	27.67	39.15
Deposit Mark-up Q1	28.42	18.32	27.17	22.26	37.00
Deposit Mark-up Q5	33.00	17.10	32.54	27.12	35.63

Figure A.11: Credit and deposit mark-ups and return on equity



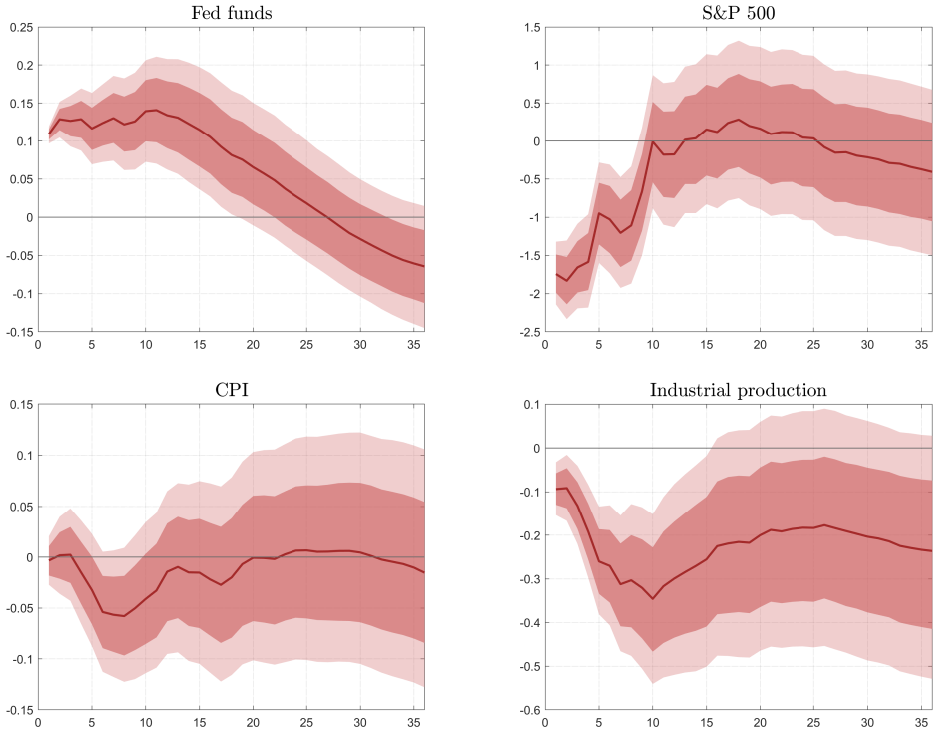
(a) Credit mark-ups

(b) Deposit mark-ups

Note. We control for time fixed effects as well as assets.

A.3 Response of macro variables

Figure A.12: Response of standard macro variables to a monetary shock



A.4 Data construction

In this section, we describe our data construction procedure. We obtain quarterly Call Report data for the period 1985q1-2020q1. For each observation in our sample, we compute book leverage as the ratio of total assets over total equity. We also compute credit and deposit mark-ups following the procedure described in Appendix A.5. Table A.3 describes in detail how we define our variables. Throughout our analysis, we truncate leverage, credit mark-ups and deposit mark-ups at the 2% and 98% level, while we truncate net income at the 1% level.[1]

For each quarter, we then compute aggregate assets, equity, and net income. That is, for each of these variables we compute the total across all observations within a given quarter. In addition, we define total leverage as the ratio of total assets over total net worth. For assets and net worth, we also compute unweighted standard deviation and statistical skewness, as well as the Herfindahl-Hirschman index (HHI).[2] For net income and leverage, we compute asset-weighted standard deviation and skewness. For credit (deposit) mark-ups, we compute asset (deposit) weighted mean, standard deviation, and skewness.

As suggested by Figure 5, there is an heterogeneous response of bank size to monetary shocks. Because of this, weighting observations by their contemporaneous asset holdings may potentially raise endogeneity concerns. As a result, we weight observations based on their average asset (deposit) holdings over the last four quarters. That is, the weight of bank i in period t is given by:

$$\omega_{i,t} = \frac{1/4 \sum_{j=1}^{4} x_{i,t-j}}{\sum_{n=1}^{N_t} 1/4 \sum_{j=1}^{4} x_{i,t-j}}$$

Where $x_{i,t}$ denotes either asset or deposit holdings. Accordingly, the way we rank observations into size quintiles is consistent with this weighting scheme. Specifically, for each period in our sample, we rank institutions based on their average asset (deposit) holdings over the last four quarters. Within each quintile, we then compute total assets, net worth, leverage, and net income, as well as weighted averages of credit and deposit mark-ups.

Notice that, since our banking variables are available only at quarterly frequency, we interpolate them to monthly frequency using a "Shape-Preserving Piecewise Cubic Hermite Interpolation" as in, for example, Miranda-Agrippino and Rey (2020). We perform this interpolation procedure at the aggregate level, that is after having computed the log of the relevant moment of a given banking variable. Figures A.9 and A.10 show that interpolating does not affect the bulk of our results.

As for the standard macroeconomic variables employed in our analysis, we download the federal

[1] We do not truncate the right tail of net income, since this would imply disregarding the largest and most influential banks in our sample.

[2] We compute the HHI for variable x according to the usual formula: $HHI_t(x) = \sum_i \left(\frac{x_{it}}{x_t}\right)^2$.

funds rate, the consumer price index, and the industrial production index at monthly frequency from the St Louis Federal Reserve. Finally, we acquire data on the price of the S&P500 index from Compustat.

Table A.3: Description of Banking Variables

Variable Name	Call Report code
Assets	RCFD2170
Net worth	RCFD3210
Leverage	RCFD2170/RCFD3210
Net income	RIAD4340
Loans	RCFD2122
Deposits	RCON2200
Treasuries and agency debt	RCFDB558
Fed funds and repo assets	RCFD3365
Securities	RCFD1754+RCFD1773
Non interest income	RIAD4079
Interest income on loans	RIAD4010
Int. inc. on fed funds and repo assets	RIAD4020
Int. inc. on treasuries and agency debt	RIADB488
Service charges on domestic deposits	RIAD4080
Interest and non interest expenses	RIAD4073+RIAD4093
Interest expenses on domestic deposits	RIAD4170-RIAD4172
Int. exp. on fed funds and repo liab.	RIAD4180
Expenses on premises	RIAD4217
Non interest expenses	RIAD4093
Salaries expenses	RIAD4135

A.5 Credit and deposit mark-up estimation

We estimate credit mark-ups using the procedure proposed in Corbae and D'Erasmo (2021). First, we define the credit mark-up for bank i in period t as:

$$\mu_{i,t} = \frac{p_{i,t}}{c_{i,t}}$$

Where $p_{i,t}$ is a measure of the price that bank i charges on loans in period t, while $c_{i,t}$ represents the marginal cost that bank i must incur in order to produce an extra unit of loans.

More specifically, we construct $p_{i,t}$ as the ratio of interest income on loans and leases over total loans and leases. $c_{i,t}$ is instead defined as the sum of two objects: the ratio of interest expenses on domestic deposits and fed funds over total deposits and fed funds (which we refer to as the cost of funds) plus marginal net non-interest expenses. In turn, marginal net non-interest expenses are computed as marginal non-interest expenses minus marginal non-interest income. We now turn to the description of these two objects.

We estimate marginal non-interest expenses by means of the following trans-log panel regression:

$$\begin{aligned}\log(NIE_{i,t}) = \alpha_i + \delta_t + \beta_{l,1}\log(l_{i,t}) + \beta_{w,1}\log(w_{i,t}) + \beta_{q,1}\log(q_{i,t}) \\ + \beta_{l,2}\log(l_{i,t})^2 + \beta_{w,2}\log(w_{i,t})^2 + \beta_{q,2}\log(q_{i,t})^2 + \beta_{l,w}\log(l_{i,t})\log(w_{i,t}) \\ + \beta_{l,q}\log(l_{i,t})\log(q_{i,t}) + \beta_{w,q}\log(w_{i,t})\log(q_{i,t}) + \varepsilon_{i,t}\end{aligned} \tag{A.1}$$

Where $NIE_{i,t}$ represents non-interest expenses, α_i and δ_t are respectively bank and time fixed effects, total loans and leases are denoted by $l_{i,t}$, $w_{i,t}$ is staff expenses,[3] and $q_{i,t}$ denotes total holdings of securities. See Table A.3 for more details on how we map variables to the Call Report data.

From Equation (A.1) it is straightforward to obtain marginal non-interest expenses as the derivative of non-interest expenses with respect to loans:

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

The estimation of marginal non-interest income relies on the exact same procedure, with the caveats that we do not include inputs (i.e. salaries) in the right hand side of the equation and that the left hand side variable is now represented by the logarithm of non-interest income, rather than non-interest expenses.

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

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

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

[3] We compute staff expenses as the ratio of salaries over assets, as in Corbae and D'Erasmo (2021).

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

$$\text{MNNIE}_{i,t} = \text{MNIE}_{i,t} - \text{MNII}_{i,t}$$

We estimate deposit mark-ups by extending the strategy proposed for credit mark-ups by Corbae and D'Erasmo (2021) and described above.

First, we define the deposit mark-up for bank i in period t analogously to before, as:

$$\mu_{i,t} = \frac{p_{i,t}}{c_{i,t}}$$

Here, $p_{i,t}$ is a proxy of the "safe revenue" that bank i is able to derive from its funds in period t, while $c_{i,t}$ represents the marginal cost that bank i must incur in order to raise an extra unit of deposits.

More specifically, we construct $p_{i,t}$ as the ratio of interest income from fed funds, US treasuries, and agency debt holdings over total fed funds, US treasuries, and agency debt holdings. As before, $c_{i,t}$ is instead 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. In turn, we compute marginal net non-interest expenses as marginal non-interest expenses minus marginal non-interest income. We now turn to the description of these two objects.

We estimate marginal non-interest expenses by means of a trans-log panel regression very similar in nature to that used for credit mark-ups:

$$\begin{aligned}\log(\text{NIE}_{i,t}) = {} & \alpha_i + \delta_t + \beta_{l,1}\log(l_{i,t}) + \beta_{w,1}\log(w_{i,t}) + \beta_{q,1}\log(q_{i,t}) + \beta_{d,1}\log(d_{i,t}) \\ & + \beta_{l,2}\log(l_{i,t})^2 + \beta_{w,2}\log(w_{i,t})^2 + \beta_{q,2}\log(q_{i,t})^2 + \beta_{d,2}\log(d_{i,t})^2 \\ & + \beta_{l,w}\log(l_{i,t})\log(w_{i,t}) + \beta_{l,q}\log(l_{i,t})\log(q_{i,t}) + \beta_{w,q}\log(w_{i,t})\log(q_{i,t}) \\ & + \beta_{l,d}\log(l_{i,t})\log(d_{i,t}) + \beta_{w,d}\log(w_{i,t})\log(d_{i,t}) + \beta_{q,d}\log(q_{i,t})\log(d_{i,t}) + \varepsilon_{i,t}\end{aligned} \tag{A.2}$$

Where $d_{i,t}$ denotes total domestic deposits,[4] while the definition of all other variables is the same as in Equation (A.1).

From Equation (A.2) it is straightforward to obtain marginal non-interest expenses as the derivative of non-interest expenses with respect to deposits:

$$\text{MNIE}_{i,t} = \frac{\partial \text{NIE}_{i,t}}{\partial d_{i,t}} = \frac{\text{NIE}_{i,t}}{d_{i,t}}\left[\beta_{d,1} + 2\beta_{d,2}\log(d_{i,t}) + \beta_{l,d}\log(l_{i,t}) + \beta_{w,d}\log(w_{i,t}) + \beta_{q,d}\log(q_{i,t})\right]$$

[4]Notice that we are not the first to include deposits together with loans as a proxy for bank output in the translog cost function, see for example Fries and Taci (2005)

The estimation of marginal non-interest income relies on the exact same procedure, with the caveats that we do not include inputs (i.e. salaries) in the right hand side of the equation and that the left hand side variable is now represented by the logarithm of non-interest income, rather than non-interest expenses.

$$\begin{aligned}\log(\mathrm{NII}_{i,t}) = {} & \alpha_i + \delta_t + \beta_{l,1}\log(l_{i,t}) + \beta_{q,1}\log(q_{i,t}) + \beta_{d,1}\log(d_{i,t}) \\ & + \beta_{l,2}\log(l_{i,t})^2 + \beta_{q,2}\log(q_{i,t})^2 + \beta_{d,2}\log(d_{i,t})^2 \\ & + \beta_{l,q}\log(l_{i,t})\log(q_{i,t}) + \beta_{l,d}\log(l_{i,t})\log(d_{i,t}) + \beta_{q,d}\log(q_{i,t})\log(d_{i,t}) + \varepsilon_{i,t}\end{aligned}$$

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

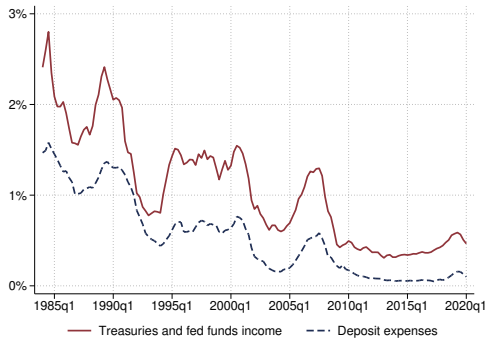
$$\mathrm{MNII}_{i,t} = \frac{\partial \mathrm{NII}_{i,t}}{\partial d_{i,t}} = \frac{\mathrm{NII}_{i,t}}{d_{i,t}}\left[\beta_{d,1} + 2\beta_{d,2}\log(d_{i,t}) + \beta_{l,d}\log(l_{i,t}) + \beta_{q,d}\log(q_{i,t})\right]$$

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

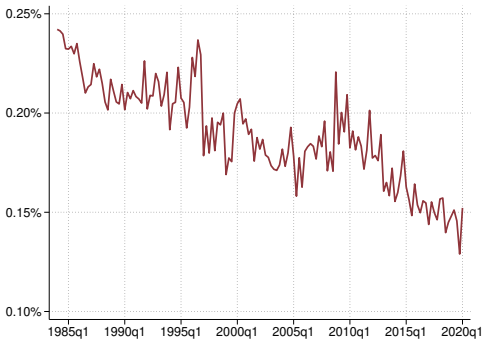
$$\mathrm{MNNIE}_{i,t} = \mathrm{MNIE}_{i,t} - \mathrm{MNII}_{i,t}$$

In Figure A.13, we perform a decomposition of our estimated deposit mark-up into its primitive components. As already described in Section 2.3, we can see a stark downward trend in both the safe rate of return and the deposit rate, with a contemporaneous reduction of the deposit spread, i.e. the wedge between these two objects. At the same time, the downward trend in marginal net non-interest expenses more than compensates for the reduction in the deposit spread, resulting into a rising trend for the deposit mark-up.

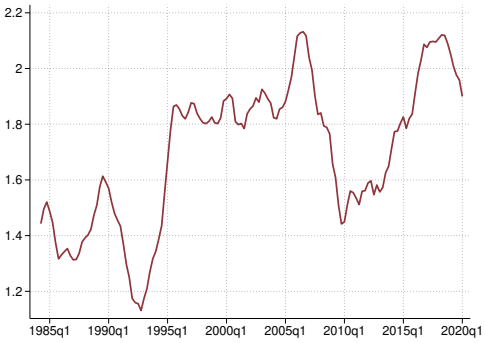
Figure A.13: Decomposition of the deposit mark-up



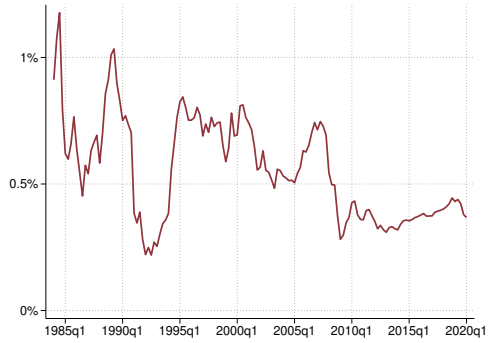
(a) Safe income and deposit expenses



(b) Marginal net non-interest expenses



(c) Deposit mark-up



(d) Deposit spread

Note: all series are deposit-weighted averages.

B Model appendix

B.1 Derivation of the household problem

$$\max_{\{C_t, B_t, b_{jt}, M_t\}} \sum_{t=1}^{\infty} \beta^t \frac{C_t^{1-\psi}}{1-\psi} + \frac{B_t^{1-\nu}}{1-\nu}$$

s.t. the budget constraint:

$$C_t + \int_0^1 b_{j,t} dj + M_t \leq R_t M_{t-1} + \int_0^1 R_{j,t}^b b_{j,t-1} + Div_t$$

and s.t. the Kimball aggregator for deposits:

$$\int_0^1 \Upsilon\left(\frac{b_{j,t}}{B_t}\right) dj = 1$$

FOC_{M_t} :

$$\beta \left(\frac{C_{t+1}}{C_t}\right)^{-\psi} = (R_{t+1})^{-1}$$

denote λ_k the Lagrange multiplier of the Kimball aggregator constraint.
FOC_{B_t} :

$$B_t^{-\nu} = \lambda_k \frac{1}{B_t} \int_0^1 \Upsilon'\left(\frac{b_{j,t}}{B_t}\right) \frac{b_{j,t}}{B_t} dj$$

$FOC_{b_{jt}}$:

$$C_t^{-\psi} - \beta R_{j,t+1}^b C_{t+1}^{-\psi} = \lambda_k \Upsilon'\left(\frac{b_{j,t}}{B_t}\right) \frac{1}{B_t}$$

Take FOC_{B_t} and solve for the Lagrange multiplier:

$$\lambda_k = \frac{B_t^{1-\nu}}{\int_0^1 \Upsilon'\left(\frac{b_{jt}}{B_t}\right) \frac{b_{j,t}}{B_t} dj}$$

Plug into $FOC_{b_{jt}}$, substitute in the FOC_{M_t} and simplify:

$$R_{j,t+1}^b = R_{t+1} - R_{t+1} \left[\frac{C_t^{\psi}}{B_t^{\nu}} \frac{\Upsilon'\left(\frac{b_{j,t}}{B_t}\right)}{\int_0^1 \Upsilon'\left(\frac{b_{j,t}}{B_t}\right) \frac{b_{j,t}}{B_t} dj} \right]$$

B.2 Klenow-Willis specification

Credit Market

Aggregator

$$\Phi\left(\frac{k}{K}\right) = 1 + (\theta_k - 1)\exp\left(\frac{1}{\epsilon_k}\right)\epsilon_k^{\frac{\theta_k}{\epsilon_k}-1}\left[\Gamma\left(\frac{\theta_k}{\epsilon_k}, \frac{1}{\epsilon_k}\right) + \Gamma\left(\frac{\theta_k}{\epsilon_k}, \frac{\left(\frac{k}{K}\right)^{\epsilon_k/\theta_k}}{\epsilon_k}\right)\right]$$

First derivative, also the inverse demand function

$$\Phi'\left(\frac{k}{K}\right) = \frac{\theta_k - 1}{\theta_k}\exp\left(\frac{1 - \left(\frac{k}{K}\right)^{\epsilon_k/\theta_k}}{\epsilon_k}\right)$$

Inverse of the first derivative, gives back relative size

$$\Psi(x) := \left(\Phi'\right)^{-1}\left(\frac{k}{K}\right) = \left(1 + \epsilon_k \log\left(\frac{\theta_k - 1}{\theta_k \Phi'\left(\frac{k}{K}\right)}\right)\right)^{\frac{\theta_k}{\epsilon_k}}$$

Deposit Market

Aggregator

$$\Upsilon\left(\frac{b}{B}\right) = 1 + (\theta_b - 1)\exp\left(\frac{1}{\epsilon_b}\right)\epsilon_b^{\frac{\theta_b}{\epsilon_b}-1}\left[\Gamma\left(\frac{\theta_b}{\epsilon_b}, \frac{1}{\epsilon_b}\right) + \Gamma\left(\frac{\theta_b}{\epsilon_b}, \frac{-\left(\frac{b}{B}\right)^{\epsilon_b/\theta_b}}{\epsilon_b}\right)\right]$$

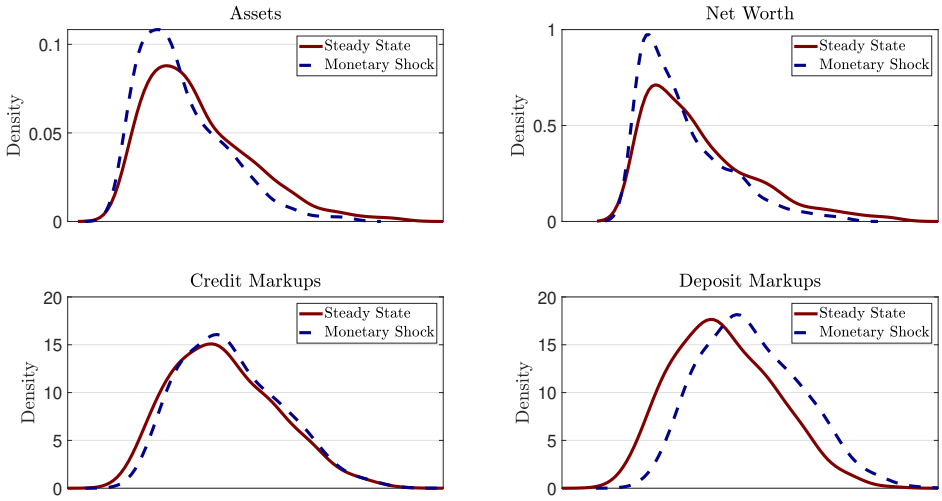
First derivative, also the inverse supply function

$$\Upsilon'(\frac{b}{B}) = \frac{\theta_b - 1}{\theta_b}\exp\left(\frac{1 + \left(\frac{b}{B}\right)^{\epsilon_b/\theta_b}}{\epsilon_b}\right)$$

Inverse of first derivative, gives back relative size

$$\Psi_b(x) := \left(\Upsilon'\right)^{-1}\left(\frac{b}{B}\right) = \left[-\left(1 + \epsilon_b \log\left(\frac{\theta_b - 1}{\theta_b \Upsilon'\left(\frac{b}{B}\right)}\right)\right)\right]^{\frac{\theta_b}{\epsilon_b}}$$

Figure B.1: Density Shifts



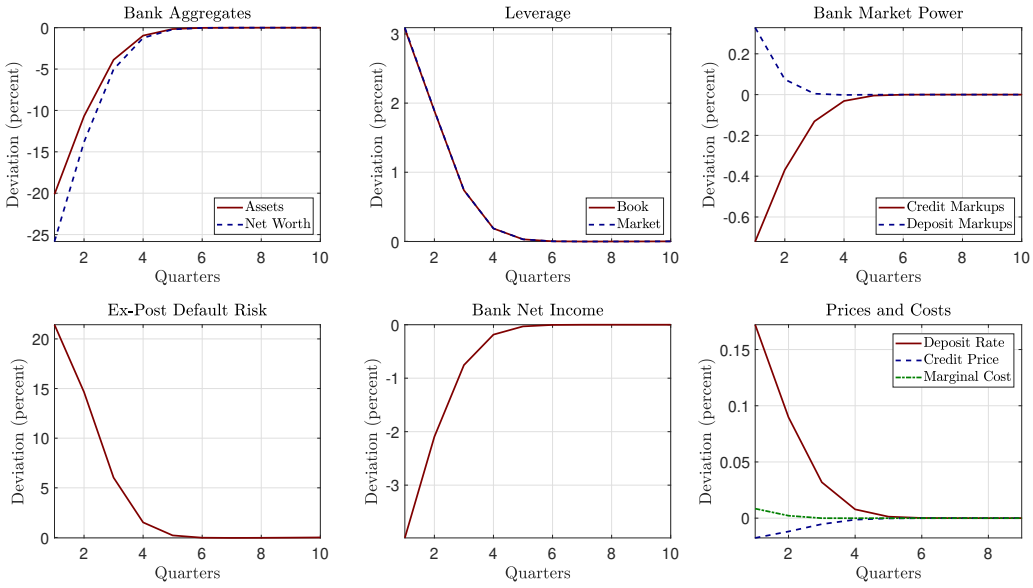
Notes: Kernel density shifts in response to a 50bps monetary surprise. Dashed blue lines plot impact densities as of period t=1, i.e. immediately after the shock. Solid red lines plot densities in the pre-shock stationary steady state.

B.3 Additional model results

We now provide auxiliary results that supplement the main text. First, in Figure B.1 we show how kernel densities of key financial variables shift following the monetary shock. When it comes to quantities, the leftward shift of both asset and net worth densities is consistent with the distribution of bank-level responses from Figure 11. The same can be said of the rightward shift of the deposit mark-up distribution. The response of credit mark-ups is once again more nuanced. Recall that, based on our findings in Figures 11 and 12 the compression of the distribution of *relative assets* causes large banks to lower and small banks to raise credit mark-ups. Also note that the latter is of greater magnitude than the former. As a result, after the monetary policy shock, large banks move towards the left of the new credit mark-up density but by a smaller distance than small banks go to the right of the density. Thus, it appears that credit mark-ups increase for every bank in the distribution. However, this is not at all the case due to the reallocation of credit market power that is happening in the background.

For completeness with our analysis of quintile-based responses, we also plot the transitional dynamics of second and third moments of key banking variables. We proxy the second (third) moment with the standard deviation (skewness) of the respective underlying distributions. For quantities such as assets and net worth we also calculate the Herfindahl index of concentration (HHI). Figures B.2 and B.3 show the results. We emphasize several noteworthy observations. When benchmarked against our empirical findings in Section A.1, the model does a very good job

Figure B.2: Second Moments

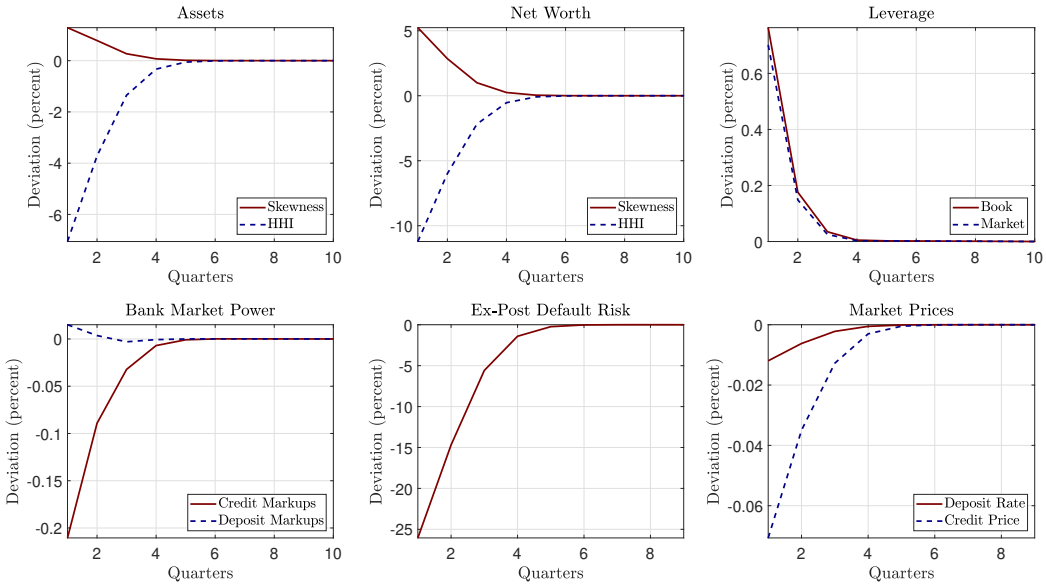


Notes: Changes in standard deviations of time-varying cross-sectional distributions of respective financial variables in response to a 50bps Taylor rule surprise.

of matching the responses of higher-order moments. For quantities such as assets and net worth, we see that while dispersion falls following the monetary contraction, skewness (HHI) increases (falls). These are the patterns that we also observe in the data. The negative effect on concentration (HHI) is particularly interesting, suggesting that expansionary policy may potentially contribute to rising banking concentration (Corbae and D'Erasmo, 2020). For market power variables, again in a data-consistent manner, we find that dispersion of credit (deposit) mark-ups falls (rises). If treated as a proxy for misallocation or inefficiency, we find that monetary policy has contrasting effects on misallocation of bank credit and liabilities. Properly accounting for *two-sided* market power is essential for capturing this channel.

In the main text, and more specifically in Figure 14 we identify the impact that credit and deposit market power have on monetary transmission individually. Now, we shut down both channels simultaneously. We compare the responses across three economies: perfect competition in both markets, CES in both markets, and the Kimball system (which is our baseline) in both markets. Figure B.4 plots the result. Because in the CES economy all mark-ups are homogenous and time-invariant, the aggregate response of either output or inflation is not distinguishable from that of the perfect competition case. However, introducing the Kimball aggregator and heterogeneous credit demand and deposit supply elasticities instead of the Dixit-Stiglitz aggregator makes a noticeable difference. Specifically, the macroeconomic response is considerably dampened. This observation

Figure B.3: Third Moments

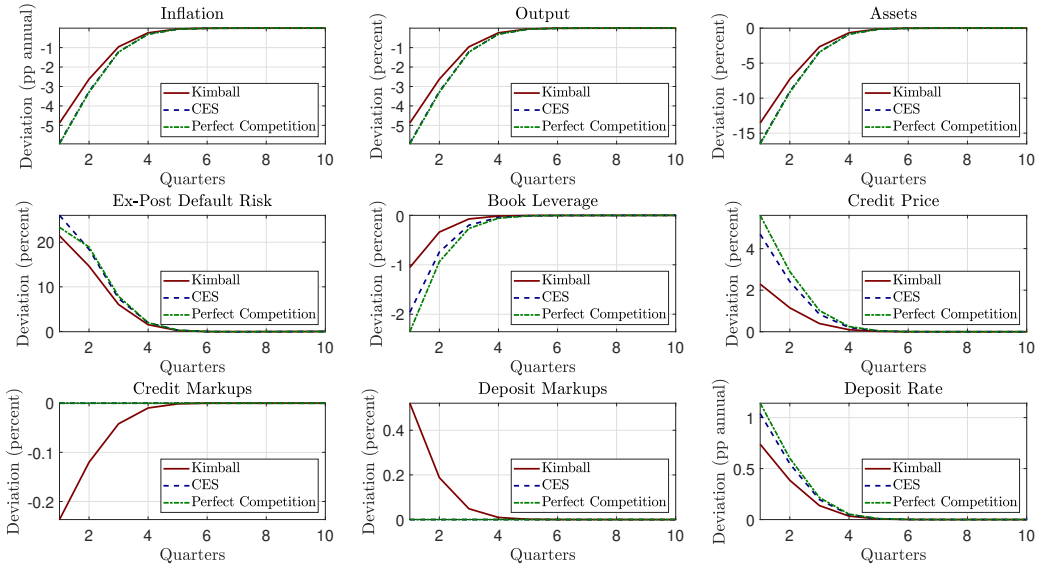


Notes: Changes in statistical skewness and the Herfindahl-Hirschman index of market concentration of time-varying cross-sectional distributions of respective financial variables in response to a 50bps Taylor rule surprise.

is best understood by recalling the individual effects of credit- and deposit-market power. The latter, by itself, is dampening monetary policy transmission while the former amplifies it. Conditional on our calibration, the dampening effect of deposit market power strongly dominates. The flexibility of our modelling approach makes it easy to calibrate the framework to different markets, periods, and countries; the relative strength of the two market power channels could change depending on the set-up.

We conclude this section by presenting heterogeneous, percentile-based deviations in response to an *expansionary* monetary shock. This supplements Figure 16 in main text which plots only the aggregate asymmetry. Figure B.5 depicts the results. The most noteworthy observation is the wide heterogeneity in the credit price responses. Large banks' credit prices are immensely rigid, exhibiting practically no response at all. This occurs because (a) credit price pass-through is declining with size in general and (b) because of the Kimball "demand kink" credit prices are *especially* more rigid when being adjusted downwards. Because expansionary shocks cause a very rigid response of prices in the top size quintile, quantities of the same large banks react by more. As a result, we get the kind of aggregate asymmetry that is seen in Figure 16: monetary expansions are considerably more impactful on both aggregate output and inflation.

Figure B.4: The Role of Market Structure Aggregators



B.4 Solution algorithms

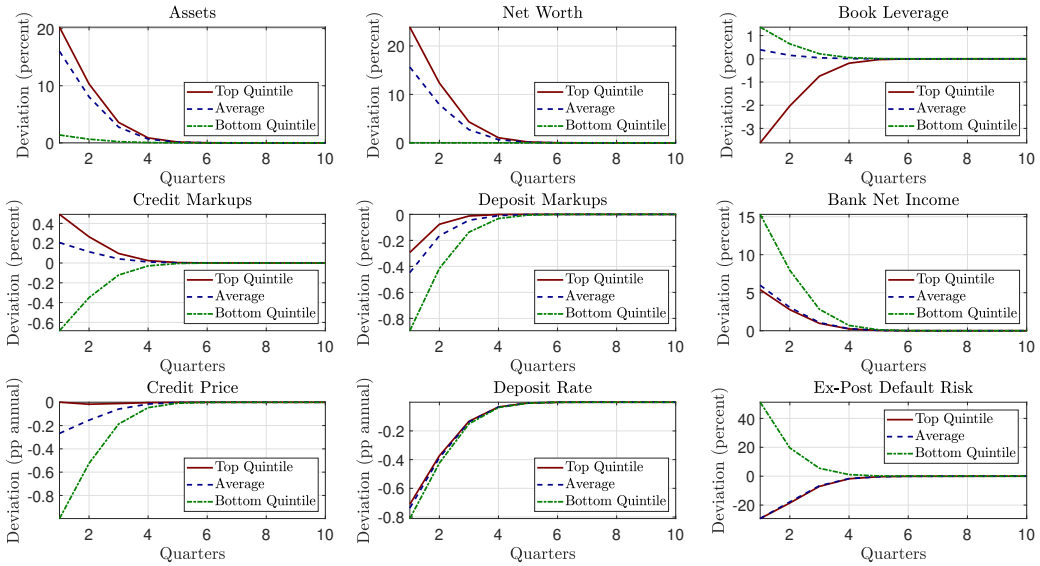
Stationary Equilibrium We solve our model non-linearly using a combination of several methods. The solution strategy consists of two general steps. First, we solve for the stationary equilibrium without aggregate uncertainty or monetary shocks, i.e. the nominal rate is equal to the target rate $\bar{R}$. Second, we solve for the transitional equilibrium in response to an unexpected "MIT" surprise shock to the Taylor rule.

1. Initialize the outer loop by guessing aggregate capital stock K. Compute final output Y and consumption C.

2. Begin by solving the household problem conditional on real wages w and the risk-free rate R. Store the resulting stochastic discount factor Λ. We use the endogenous gridpoint method of Carroll (2006) for speed, although any standard method such as value function or policy function iteration suffices.

3. Conditional on Λ and C and initial guesses for the aggregate state vector we solve the banking problem with projection methods. For the idiosyncratic state vector, we use ten gridpoints for net worth. We assume throughout that there are three permanent profitability types whose κ_j correspond to three nodes of the Tauchen-Hussey quadrature of a unit-root process with the error drawn from $\mathcal{N}(0, \sigma_\kappa)$. We discretize ξ_j with 7 nodes. We deal with the occassionally binding leverage constraint the following way. On every grid point we first assume that the

Figure B.5: Asymmetric Responses by Percentile



Responses by quintiles (20%) of the steady-state bank net worth distributions to a 50bps negative Taylor rule surprise.

constraint binds. Under this assumption, we back out the Lagrange multiplier and check if the assumption holds. If the constraint is in fact slack, we re-solve the problem using the Chris Sims' global minimization routine. In both stages, it's essential that banks internalize the impact of private quantity decisions onto prices via the Kimball credit demand and deposit supply functions.

4. Run a long stochastic simulation, for each permanent profitability type, using the newly computed policy functions. Compute aggregate capital K' as an unweighted average of bank-level capital choices.

5. Conditional on the newly computed K', solve the New Keynesian block and determine the rate of final good inflation. In the steady state, gross inflation will always equal to unity.

6. Calculate the percentage difference between K and K'. Update the guesses and return to step 2. Proceed upon convergence.

Monetary Policy Shocks: Transitional Equilibrium Our approach is a variant of the well-known shooting algorithm. Our model is in discrete time and we build upon the basic algorithm that is transparently laid out in Boppart et al. (2018). For continuous-time frameworks, a similar approach is described in Kaplan et al. (2018).

1. Choose time T at which it is conjectured that the economy is back in the steady state equilibrium.

2. Project the mean-reverting path of the MIT shock to monetary policy $\{\epsilon_{mt}\}_{t=1}^{T}$ that hits the economy at time t=1.

3. Guess a path for the aggregate capital stock $\{K_t\}_{t=1}^{T}$. Compute production, consumption, and nominal wages along the transition path. Guess a path of bank franchise values $\{V_t\}_{t=1}^{T}$ and policy functions for net worth $\{n_t'\}_{t=1}^{T}$.

4. Solve the New Keynesian block backwards from t = T – 1 . . . 1 by setting $\Pi^T = \Pi^{SS}$.

5. Solve the household problem backwards from t = T – 1 . . . 1 by setting the policy function for deposits to its steady-state value $b'^T = b^{SS}$. Store the implied path of $\{\Lambda_t\}_{t=1}^{T}$.

6. Solve the banking problem backwards conditional on the paths of endogenous state variables. Within-period solution follows the same steps as in the stationary case.

7. Simulate the distribution of banks forward using the just-computed policy functions.

8. Compute the new path of $\{K_t'\}_{t=1}^{T}$ from the time-varying cross-sectional distribution.

9. Calculate the percentage difference between $\{K_t'\}_{t=1}^{T}$ and the old candidate $\{K_t\}_{t=1}^{T}$.

10. If the maximal difference is below the tolerance level, stop the algorithm. Otherwise, update $\{K_t\}_{t=1}^{T}$ very slowly and revert back to Step 3.

References

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

Carroll, C. (2006): "The method of endogenous gridpoints for solving dynamic stochastic optimization problems," *Economics Letters*, 91(3).

Corbae, D. and P. D'Erasmo (2020): "Rising bank concentration," *Journal of Economic Dynamics and Control*, 115, 103877.

——— (2021): "Capital Buffers in a Quantitative Model of Banking Industry Dynamics," *Econometrica*, 89, 2975–3023.

Fries, S. and A. Taci (2005): "Cost efficiency of banks in transition: Evidence from 289 banks in 15 post-communist countries," *Journal of Banking Finance*, 29, 55–81, banking and the Financial Sector in Transition and Emerging Market Economies.

Jarociński, M. and P. Karadi (2020): "Deconstructing Monetary Policy Surprises - The Role of Information Shocks," *American Economic Journal: Macroeconomics*, 12, 1–43.

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

Miranda-Agrippino, S. and H. Rey (2020): "U.S. Monetary Policy and the Global Financial Cycle," *The Review of Economic Studies*, 87, 2754–2776.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business Economic Statistics*, 20, 518–529.