Federal Deposit Insurance Corporation • Center for Financial Research

Working Paper Series

Bank Concentration and Monetary Policy Pass-Through

Isabel Gödl-Hanisch
LMU Munich

June 2022

FDIC CFR WP 2022-06

fdic.gov/cfr

Bank Concentration and Monetary Policy Pass-Through

Isabel Gödl-Hanisch*

LMU Munich

This version: March, 2022

Click here for the latest version

Abstract

This paper analyzes the implications of the recent rise in bank concentration for the transmission of monetary policy. First, I use branch-level data on deposit and loan rates to evaluate the monetary policy pass-through conditional on the level of local bank concentration and bank capitalization. I find that banks operating in high-concentration markets and under-capitalized banks adjust short-term lending rates more, particularly when the policy rate increases. Second, I build a theoretical model with heterogeneous banks that rationalizes the empirical findings and explains the underlying mechanism. In the model, monopolistic competition in local deposit and loan markets along with bank capital requirements impose frictions on the pass-through to the real economy. Counterfactual analyses highlight that the rise in bank concentration strengthens monetary policy pass-through by two channels: the market power and capital allocation channel. Both channels further enhance monetary policy transmission to output and investment, amplify the credit cycle, and flatten the Phillips curve.

Keywords: monetary transmission, bank heterogeneity, monopolistic competition, bank regulation.

JEL codes: E44, E51, E52, G21.

---

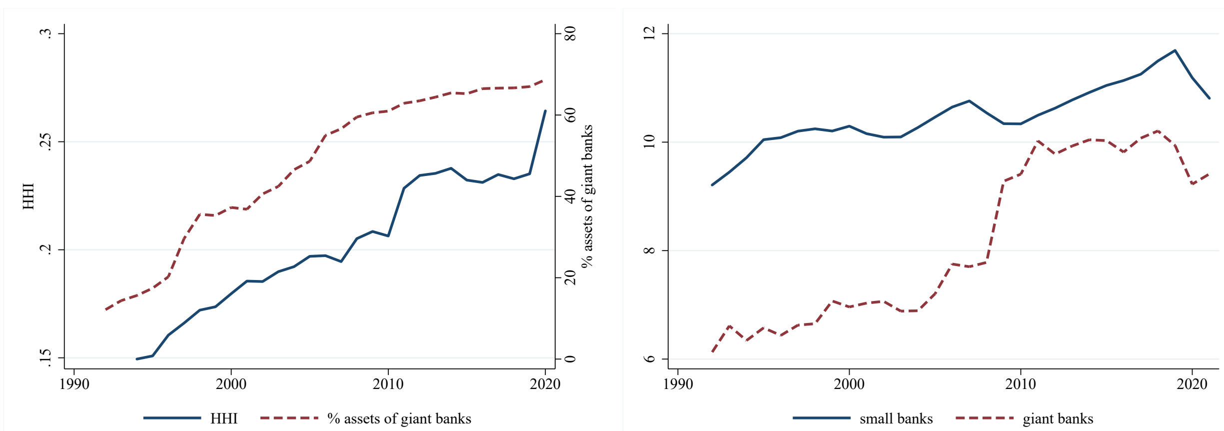
*Correspondence: isabel.goedl-hanisch@econ.lmu.de. I am grateful to Eric Sims, Christiane Baumeister, Cynthia Wu, and Rüdiger Bachmann for their guidance and support, and to Rosalind Bennett, Christopher Martin, and Jonathan Pogach for their insightful comments and feedback. I received helpful comments from seminar participants at the Federal Reserve Board, European Central Bank, 2020 CEBRA annual meeting, Central Bank of Chile, 15^{th} Economics Graduate Student Conference at Washington University in St. Louis, 2^{nd} Workshop for Women in Macroeconomics, Finance and Economic History, University of Würzburg, and 20^{th} FDIC/JFSR Bank Research Conference. I thank the Notre Dame Department of Economics and Mendoza School of Business for purchasing the microdata for this project and the ISLA Graduate Student Research Award for financial support.

1 Introduction

Over the last two decades, the U.S. banking sector has become increasingly concentrated, as relaxed banking regulation before the financial crisis and bank consolidation after the financial crisis significantly reduced the number of banks in many local banking markets.[1] In 1994, the largest U.S. banks[2] owned 16% of total commercial bank assets; that share increased to 69% by 2020. During the same time, the local Herfindahl-Hirschman Index (HHI) steadily grew from a moderate level of 0.15 in 1994 to a highly concentrated level of 0.26 in 2020, as shown in Panel (a) of Figure 1.[3] This paper studies the question of how and whether the recent rise in bank concentration has altered monetary policy transmission to the real economy.

Figure 1: The U.S. banking sector over time

(a) Bank concentration (b) Core capital ratio



Notes: HHI is shown at the average county level and is weighted by total deposits, % assets of giant banks is the asset share of banks $>$ \$100 billion in assets (in \$2018), and core capital ratio measures mean core capital over risk-weighted assets by group. Source: Federal Deposit Insurance Corporation.

To assess the role of bank concentration for monetary policy pass-through, it is crucial to look at observed differences in retail rates and lending volumes within a given bank across regions as well as across bank institutions within a region. The variation in retail rates serves to shed light on how the composition of local markets and the size distribution of banks affect the aggregate transmission of monetary policy via two channels. The first channel is the *market power channel*: a higher concentration in local banking markets leads

[1] For example, the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994 permitted banks to open branches across states, and the Glass-Steagall Act's repeal in 1999 allowed commercial banks to offer both securities and insurance (Corbae and D'Erasmo, 2020).

[2] Giant banks are defined as those above \$100 billion in assets (in \$2018). Classification and cutoff follow the definition of the Federal Reserve Bank for large financial institutions.

[3] Appendix A.1 decomposes national bank concentration growth and finds within-county growth and rising concentrations in counties with deposit inflows contribute significantly to the overall effect (Figure A.1).

to a widening wedge between the central bank's policy rate and the commercial banks' loan and deposit rates. The second channel is the *capital allocation channel*: a higher banking concentration implies that giant banks, which tend to have relatively low capital ratios, as shown in Panel (b) of Figure 1, handle an increasing share of total loans and deposits. This might be expected to amplify financial frictions arising from regulatory requirements on giant banks. In the past years, Basel III reforms had the effect of mitigating the decline in capitalization that was driven by increasing concentration.

Extant literature has largely neglected the effects of the banking sector's composition on monetary policy transmission. While research has shown bank market power (e.g., Drechsler et al., 2017; Scharfstein and Sunderam, 2016) and bank size and capitalization (e.g., Kashyap and Stein, 2000; Van den Heuvel, 2002) both impact the effectiveness of monetary policy in isolation, there is little evidence on the relative importance of each channel. Nor have researchers provided compelling evidence about the channels' combined implications for monetary policy transmission. The contribution of this paper is to emphasize the importance of compositional effects for the transmission of monetary policy and to demonstrate that a partial analysis falls short of accounting for interaction effects and thus may lead to inaccurate conclusions.

This paper starts by building a simple model of heterogeneous monetary policy pass-through to retail rates inspired by the canonical Monti–Klein model.[4] To micro-found the differences between branches of the same bank across locations and the differences across bank institutions in the same location, I combine two conventional building blocks. First, banks hold market power in local deposit and loan markets. Second, banks face a capital requirement that imposes additional friction on monetary policy pass-through. The theoretical model predicts that monetary policy pass-through to loan rates is an increasing function of local bank concentration, as the markup is a multiplier on the policy rate; whereas monetary policy pass-through to deposit rates is a decreasing function of local bank concentration, as the markdown is a multiplier on the policy rate. The model also predicts that monetary policy pass-through to loan rates is a decreasing function of bank capitalization, as the capital constraint imposes an additional lending cost. Further, it suggests interaction effects between a bank's capitalization and market power.

In the empirical part of the paper, I first present novel facts on rate dispersion and cyclical spreads using confidential U.S. bank branch-level data from *RateWatch* from January 1998 to March 2019. I document substantial rate dispersion within banks and locations, counter-cyclical loan spreads and rate dispersion, and asymmetric adjustment in line with the assumptions of the theoretical model. I then test the model's predictions

[4]See Monti (1972) and Klein (1971).

by studying monetary policy pass-through to consumer retail rates. I define *monetary policy pass-through* as the extent to which loan and deposit rates respond to changes in the monetary policy rate. To control for potential endogeneity in monetary policy, I use monetary policy surprises from Nakamura and Steinsson (2018) as instruments for the policy rate. Using state-dependent local projections, I also allow for asymmetries between periods of monetary tightening and easing. To assess the relative importance of the *market power channel* and the *capital allocation channel*, I exploit variation in local bank concentration and bank capitalization. The empirical results confirm the model's predictions. Monetary policy pass-through to loan rates is higher (i) for branches operating in high-concentration counties, (ii) for banks with low capital ratios, and during periods of (iii) low system-wide capitalization, and (iv) monetary tightening versus easing.

To quantify the relative importance of the different frictions and perform counterfactual analyses, I embed the simple model into a dynamic New Keynesian model, similar to Gerali et al. (2010). With segmented markets, patient households provide deposits to the banking sector, while impatient households and entrepreneurs demand credit. The introduction of financial frictions on the banking side impairs the intermediation of credit between the agents. In addition to the simple model, I assume that banks are subject to asymmetric costs when adjusting loan supply due to increasing operating costs during periods of low interest rates and high demand. Asymmetric bank lending adjustment costs therefore lead to an incomplete pass-through, consistent with the downward stickiness observed in the data. For the counterfactual analyses, I extend the model to heterogeneous bank headquarters facing size-dependent capital requirements and branches operating in spatially segmented markets with differing bank concentrations.

The counterfactual analyses show that increasing bank concentration from 1994 to 2019 amplified monetary policy pass-through to loan rates. In other words, loan rates and bank lending became more sensitive to monetary policy changes. Decomposing the total pass-through change over time reveals that the *market power channel*, increasing markups, and local market share changes are the most significant contributors to the overall effect. The impacts of the *capital allocation channel*, rising capital requirements, and giant banks' market share changes over time are relatively small. However, additional significant interaction effects emerge as higher market power increases the response to changes in marginal costs and financial frictions, a part underestimated in a partial analysis as in most literature. Another insight is that the extent of macroeconomic implications depends on whether the households and firms are financially constrained. Adding borrowing constraints à la Iacoviello (2005) to households and firms lowers their sensitivity to loan rates, and compositional shifts in the banking sector become less important.

Further, rising bank concentration alters monetary policy transmission to the macroeconomy. It amplifies the monetary transmission to output and investment but dampens its impact on inflation. The opposing effects on output and inflation lead to a flatter observed empirical Phillips curve over time, consistent with recent U.S. data (Ball and Mazumder, 2011; Hazell et al., 2020; Kuttner and Robinson, 2010; Matheson and Stavrev, 2013). There are two sets of factors at play in the background. First, the slope of the Phillips curve depends on the level of resource costs from the banking sector, leading to a wealth effect. Rising bank concentration increases these costs and widens the gap between production and effective output, breaking the close link between output and marginal costs. Second, labor supply frictions, specifically wage rigidity and habit formation, individually and jointly lead to a further decoupling of output, marginal costs, and inflation and flatten the Phillips curve over time.

The remainder of this paper is structured as follows. Section 2 discusses the related literature. Section 3 proposes a simple model of heterogeneous monetary policy pass-through. Section 4 describes the data set. Section 5 presents a summary of novel stylized facts on the pass-through to deposit and loan rates. Section 6 outlines the richer theoretical model and performs counterfactual analyses, decomposes the total effect of rising bank concentration on monetary transmission, and studies the implications for the Phillips curve. Section 7 concludes.

2 Related Literature

This paper bridges research explaining differences in monetary policy pass-through based on bank characteristics and local market conditions. Similar to the structural approach of Wang et al. (2018), I quantify the implications of several frictions for monetary policy pass-through, comparing the role of loan and deposit market power and capital constraints shown to be important by Kashyap and Stein (2000), Kishan and Opiela (2000), Altavilla et al. (2019), and Van den Heuvel (2002).[5] I add to Wang et al. (2018)'s analysis of bank lending by looking at the cross-section of retail rates, taking into account that banks operate in local markets, and by offering micro-foundations for the various frictions at play.[6] Drechsler et al. (2017) establish that banks in highly concentrated markets have a

[5] Kashyap and Stein (2000) find a higher pass-through for small and less liquid banks. Kishan and Opiela (2000) study the interaction with regulatory policies, Altavilla et al. (2019) the relevance of leverage and non-performing loans, and Van den Heuvel (2002) of capital requirements.

[6] Most extant papers study the effect on total lending or impute rates from interest income data (Drechsler et al., 2018), an approach prone to composition effects, such as from shifting borrower risk.

lower pass-through to deposit rates.[7] Similarly, Scharfstein and Sunderam (2016) analyze the pass-through of mortgage-backed securities (MBS) yields to mortgage refinancing and the role of bank concentration therein, finding that banks in high-concentration markets are less sensitive to changes in MBS yields. While my paper also focuses on mortgages, the emphasis lies on the pass-through of changes in the policy rate to short-term mortgage rates and the role of bank concentration. Another contribution is to connect the findings on local bank concentration and bank characteristics. On top of that, I control for endogenous changes in the policy rate as a regressor to rule out a potential response to credit conditions.[8] Using local projections instead of panel techniques shows the pass-through dynamics and easily incorporates state-dependencies,[9] such as asymmetries between monetary easing and tightening that have been highlighted in other contexts.[10] My results are also consistent with findings on higher markups and concentration in the financial sector over time (Corbae and D'Erasmo, 2020; De Loecker et al., 2020).

On the theoretical side, I build on the canonical studies by Monti (1972) and Klein (1971). Similar to Gerali et al. (2010) and Andres and Arce (2012), I model the banking sector with monopolistic competition, which assumes that deposits and loans are baskets of differentiated products with constant elasticity of substitution leading to a constant markup. Gerali et al. (2010) compare the transmission of shocks with and without financial frictions in the banking sector in a New Keynesian model, finding that bank capital requirements, imperfect competition, and sticky rates alter monetary policy transmission. I extend their framework to include heterogeneous bank headquarters and branches to compare the pass-through in different banking environments. I also regard my results as complementary to recent work by Levieuge and Sahuc (2021) on downward loan rate rigidity that can generate similar state-dependent dynamics but falls short of micro-founding the source of adjustment asymmetries. In addition, my paper fits into the growing theoretical literature on the state-dependency of monetary policy transmission. Amongst them, Brunnermeier and Koby (2018) demonstrate that an accommodative monetary policy shock reverses and becomes contractionary when the policy rate falls below a certain level. Likewise, Wang (2019) and Ulate (2021) study monetary policy transmission to deposit and loan rates, focusing on low and negative rates. In contrast, my paper focuses on the cross-sectional pass-through of monetary tightening and easing to retail rates.

[7] There is also extensive literature deposit rates and concentration, e.g., Berger and Hannan (1989).

[8] Bluedorn et al. (2017) find more substantial heterogeneity when using Romer and Romer (2004) narrative monetary shocks compared to federal funds rate changes.

[9] Similar to Ramey and Zubairy (2018) who study state-dependent government spending multipliers.

[10] Peltzman (2000) documents asymmetric price adjustment in various industries, Borenstein et al. (1997) examines gasoline markets, and Neumark and Sharpe (1992), Yankov (2014) and Driscoll and Judson (2013) consider deposit markets.

3 Simple Model of Heterogeneous Pass-Through

To provide intuition for the empirical section, I build a simple model of heterogeneous monetary policy pass-through to retail rates inspired by the canonical Monti–Klein model.[11] The proposed model rationalizes retail rate differences between *branches* of the same bank across locations and *bank institutions* within the same location. The model makes three predictions for cross-sectional pass-through differences, ceteris paribus: (i) a higher pass-through to loan rates in high-concentration locations, (ii) a lower pass-through to deposit rates in high-concentration locations, and (iii) a higher pass-through for low capitalization banks. The model also suggests an interaction between the *market power channel* and *capital allocation channel*.

In the stylized model, banks are financial intermediaries and originate loans funded by deposits and bank capital in different locations c. Financial regulations require banks to hold specific bank capital ratios. Assume that banks are exogenously endowed with heterogeneous bank capital, implying variation in bank lending and deposit holdings across banks due to size-dependent capital constraints.[12] Banks operate under monopolistic competition, taking the local market conditions into account, wherein market power could arise from spatial and product differentiation. Table 1 shows a bank's balance sheet with loans, L_i^c, and reserves, R_i^c, as assets, and deposits, D_i^c, and bank capital, K_i^c, as liabilities.

Table 1: Bank's balance sheet

Assets		Liabilities	
Loans	L_i^c	Deposits	D_i^c
Reserves	R_i^c	Bank capital	K_i^c

Each bank i in location c is static and seeks to maximize profit, $\Pi_i^c = r_i^{l,c} L(r_i^{l,c}) + r^f R_i^c - r_i^{d,c} D(r_i^{d,c})$, subject to (i) a capital requirement, $K_i^c \geq \nu_i L_i^c$, governed by ν_i, the minimum bank capital adequacy ratio; (ii) local loan demand, $L(r_i^{l,c}) = \left(\frac{r_i^{l,c}}{\bar{r}^{l,c}}\right)^{-\epsilon^{l,c}} \bar{L}^c$, depending on local elasticity, $\epsilon^{l,c}$, aggregate loan rate, $\bar{r}^{l,c}$, aggregate loan demand, $\bar{L}^c$, and offered loan rate, $r_i^{l,c}$; (iii) local deposit supply, $D(r_i^{d,c}) = \left(\frac{r_i^{d,c}}{\bar{r}^{d,c}}\right)^{-\epsilon^{d,c}} \bar{D}^c$, depending on local elasticity, $\epsilon^{d,c}$, aggregate deposit rate, $\bar{r}^{d,c}$, aggregate deposit supply, $\bar{D}^c$, and offered deposit rate, $r_i^{d,c}$; and (iv) a balance sheet constraint, $L_i^c + R_i^c = D_i^c + K_i^c$.[13]

[11] For more details, see Freixas and Rochet (2008); Klein (1971); Monti (1972).

[12] Large banks must hold higher bank capital ratios under the Federal Reserve Bank's capital framework.

[13] A further reserve requirement would impose additional friction and affect loan and deposit rates. I abstract from a reserve requirement, as such likely has not been binding in the last years, particularly since the Federal Reserve began to pay interest on reserves in 2008. In March 2020, the Federal Reserve eliminated reserve requirements. For details, see the website of the Federal Reserve.

Solving the maximization problem and rewriting the first-order conditions yields the loan and deposit rate decision as a function of the local markup and markdown on bank i's marginal cost and policy rate, r^f, where ϕ_i reflects the multiplier on the capital constraint:

$$r_i^{l,c} = \underbrace{\frac{\epsilon^{l,c}}{(\epsilon^{l,c}-1)}}_{\text{markup}} \underbrace{\left(r^f + \nu_i \phi_i\right)}_{\text{marginal cost}}, \tag{1}$$

$$r^{d,c} = \underbrace{\frac{\epsilon^{d,c}}{(\epsilon^{d,c}-1)}}_{\text{markdown}} r^f. \tag{2}$$

As shown in equation (1), marginal costs for bank lending are heterogeneous across banks due to differences in the capital requirement, ν_i, interacting with ϕ_i, the multiplier on the capital constraint. Lending is relatively more costly for constrained banks, increasing their marginal costs and loan rates. Equation (2) indicates that the policy rate, r^f, solely influences deposit rates. The capital requirement does not have an effect. Further, loan and deposit rates depend on markups and markdowns, which vary across locations due to monopolistic competition in local markets. The markups and markdowns are functions of loan demand, $\epsilon^{l,c}$, and deposit supply elasticities, $\epsilon^{d,c}$, in location c. The lower the elasticity, the higher the markup and lower the markdown, linked to high concentration.

The total derivatives of the loan and deposit rate with respect to policy rate, r^f, inform about monetary policy pass-through:

$$\frac{dr_i^{l,c}}{dr^f} = \underbrace{\frac{\epsilon^{l,c}}{(\epsilon^{l,c}-1)}}_{\substack{\textit{market power} \\ \textit{channel}}} + \frac{\epsilon^{l,c}}{(\epsilon^{l,c}-1)} \underbrace{\nu_i \frac{d\phi_i}{dr^f}}_{\substack{\textit{capital allocation} \\ \textit{channel}}} \tag{3}$$

$$\frac{dr^{d,c}}{dr^f} = \underbrace{\frac{\epsilon^{d,c}}{(\epsilon^{d,c}-1)}}_{\substack{\textit{market power} \\ \textit{channel}}} \tag{4}$$

Equation (3) indicates that changes in the policy rate, r^f, affect loan rates by more in relatively less competitive regions. Intuitively, banks with high market power can easily pass changes in marginal costs to the consumer. Market structure shifts thus affect loan rate pass-through directly: A lower elasticity of loan demand leads to higher markups and pass-through (i.e., the *market power channel*). Further, the magnitude of the pass-through of policy rate changes to consumers is larger for banks with a low capital ratio. Hence, capital requirement shifts directly affect loan rate pass-through: Lower capitalization,

ν_i, leads to a higher pass-through (i.e., the *capital allocation channel*). The reason is that the multiplier on the constraint, ϕ_i, declines in response to a monetary tightening as higher rates curb loan demand. Increased capitalization allows banks to benefit more from an easing constraint. Conversely, this means that loan rates of more levered, less capitalized banks fluctuate more. Further, a non-negligible interaction effect results, as market power amplifies the capital allocation channel. In contrast, deposit rate pass-through, as shown in equation (4), increases with competitiveness due to a declining markdown and is unaffected by the capital constraint. The extended model in Section 6 embeds this framework and provides proofs. The empirical section tests and quantifies the cross-sectional pass-through predictions:

1. Pass-through to loan rates increases with bank market power: $\epsilon^{l,c} \downarrow \Rightarrow \frac{dr_i^{l,c}}{dr^f} \uparrow$.
2. Pass-through to loan rates declines with bank capitalization: $\nu_i \uparrow \Rightarrow \frac{dr_i^{l,c}}{dr^f} \downarrow$.
3. Pass-through to deposit rates declines with bank market power: $|\epsilon^{d,c}| \downarrow \Rightarrow \frac{dr^{d,c}}{dr^f} \downarrow$.

4 Data Description

This paper combines multiple banking data sources, county-level and national macroeconomic data, and monetary policy surprises to study pass-through to loan and deposit rates. First, I use a panel of offered deposit and loan rates at a branch level for U.S. commercial banks and credit unions from January 1998 to March 2019, provided by *RateWatch*.[14] The data provider regularly surveys 76,000 financial institution locations and collects quotes of deposits, mortgages, and consumer loan rates. The sampled loan rates provide information for the "best" borrowers, i.e., those with exceptional FICO scores,[15] for a particular constant loan volume.[16] In the case of mortgages, the volume is $175,000. *RateWatch* serves as an advertisement and informational platform for consumers and business-to-business marketers, who expect the posted rates to be accurate and available. For more information on the survey and a sample pricing sheet, see Appendix A.3. Second, using the branch identifier, the rate data is then merged with the *FDIC*'s Summary of Deposits, including annual county-level branch deposits and historical ownership information. Third, the sample is combined with the Statistics on Depository Institutions (SDI), including bank balance sheet information, using the bank identifier.

[14] *RateWatch* is part of S&P Global since 2018. The loan rate data coverage starts in January 2000.

[15] The credit score cutoff is for most banks 740 or higher, see for example Bank of America or Chase.

[16] The data set includes fixed and adjustable mortgage rates. The j-year hybrid rate (i.e., j-year ARM) is fixed for j years, then indexed to a conventional interest rate but adjusted for 30-j years.

I construct three key metrics to evaluate heterogeneous pass-through: (i) local bank concentration, (ii) bank-level characteristics, and (iii) a monetary policy measure.

Measuring local concentration. The canonical market concentration measure is the Herfindahl-Hirschman Index (HHI). The U.S. Department of Justice's antitrust division applies the measure to assess bank mergers. The HHI measures the sum of each bank institution's squared market share by county for each point in time:[17]

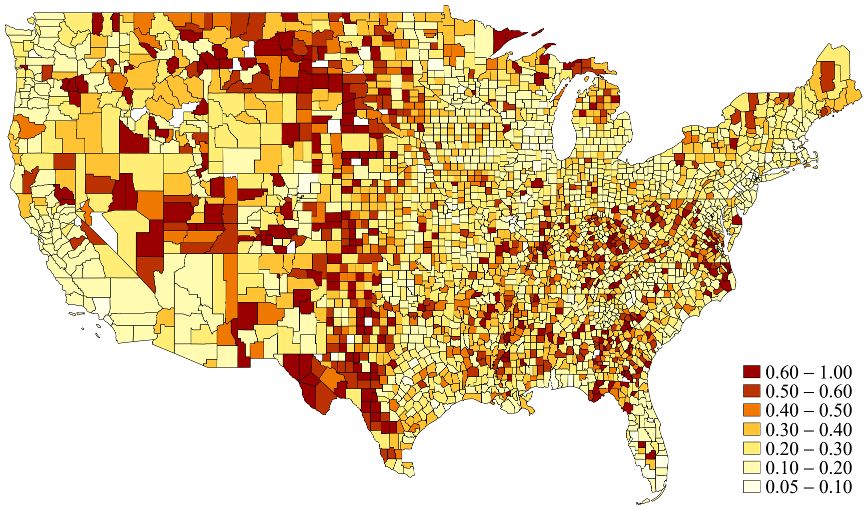
$$HHI_{c,t} = \sum_{i=1}^{I} s_{c,t,i}^2 = s_{c,t,1}^2 + s_{c,t,2}^2 + ... + s_{c,t,I}^2, \tag{5}$$

where $s_{c,t,i}$ reflects bank i's market share in county c. An HHI of 1 indicates a perfect monopoly, and $\frac{1}{I}$ is an oligopoly with I equal-sized banks. The Department of Justice classifies a market with an HHI between 0.1 and 0.18 as "moderately concentrated" and above 0.18 as "highly concentrated," according the Federal Reserve Bank of St. Louis. In the baseline, I construct the HHI by county-time and based on branch deposits per county, similar to Drechsler et al. (2017). Figure 2 shows bank concentration across counties in the US in 2019. Considerable cross-sectional variation emerges among the HHIs ranging from 0.05 to 1, both across and within states. For example, Florida's Leon County had an HHI of 0.1 in 2019, while surrounding counties Jefferson and Wakulla had HHIs of 0.66 and 0.44. Figure A.2 in Appendix A.1 provides evidence of shifts in local bank concentration between 1994 and 2019 and examines in more detail what explains the increase in bank concentration at the aggregate level. A large proportion of counties, particularly those with large banking sectors, observed increasing concentration over time.

Instead of focusing on deposit market concentration, an alternative is to look directly at mortgage market concentration similar to Scharfstein and Sunderam (2016) using Home Mortgage Disclosure Act (HMDA) data. Unlike the deposit-based measure, the mortgage-based measure pertains to flows, resulting in higher volatility and less reasonable estimates at a granular level, but includes credit unions and non-bank lenders. Despite these fundamental differences, both concentration measures are highly correlated, as shown in Appendix A.2. Deposit market concentration can therefore provide a good proxy of loan market power. Appendix A.6 contrasts the empirical results for both measures and confirms the main results' robustness to the choice of concentration measure.

[17] A market is defined at the county level, consistent with the evidence of the average distance to a lender of 1.25 miles in Canada (Allen et al., 2019) and Fannie Mae's National Housing Survey (Q1 2019) documenting that 2 of 5 recent home buyers did not shop around for mortgage lenders. The results are robust to defining competition at an MSA level instead of county level.

Figure 2: Bank concentration by county



Notes: 2019 HHI by county based on deposits. Source: FDIC Summary of Deposits.

Measuring bank capitalization. I define bank capitalization as the bank capital (equity) to total assets ratio, a key pillar of the Basel III requirements.[18] Alternatively, in line with the theoretical model, one could use the risk-weighted measure. Since the financial crisis, risk-weighted measures have become an integral part of bank regulation. At a bank level, all measures are strongly correlated. Robustness checks using the *core-capital ratio*, the *total risk-based capital ratio* and the *tier 1 risk-based capital ratio* yield similar results.

Measuring monetary policy. I measure monetary policy changes using surprises (Nakamura and Steinsson, 2018) computed from financial market variable changes within 30 minutes around Federal Open Market Committee meetings. They correspond to the first principal component of high-frequency movements in federal funds and Eurodollar futures with one year or less maturity.[19] The policy indicator captures, therefore, a forward guidance component, consistent with the short-term loan rate maturity.[20] Other monetary policy surprises, Romer and Romer (2004) narrative monetary policy shocks, and raw changes in the federal funds rate confirm the results (Appendix A.5).

[18]For details, see the website of the Bank for International Settlements (BIS) on Basel III (link).

[19]The principal component analysis includes five futures: (i) the current month, (ii) and three-month ahead federal funds, and the eurodollar at the horizons of (iii) two, (iv) three, and (v) four quarters.

[20]I replicate and extend Nakamura and Steinsson (2018)'s monetary policy surprise series up to 2019.

5 Empirical Findings

This section presents novel empirical evidence on loan and deposit rates using branch-level data from *RateWatch*, which has not been studied in the cross-section.[21] First, I examine loan and deposit spreads and rate dispersion across branches and time to assess policy rate pass-through. Second, I look closer at monetary policy pass-through to loan rates using state-dependent local projections conditioning on the level of local bank concentration and bank capitalization, and monetary tightening versus easing to explain time-varying cross-sectional dispersion. Previous research offers extensive evidence on the link between deposit rate pass-through and bank concentration (e.g., Drechsler et al., 2017); my simple model suggests that bank capitalization does not affect deposit rate pass-through.

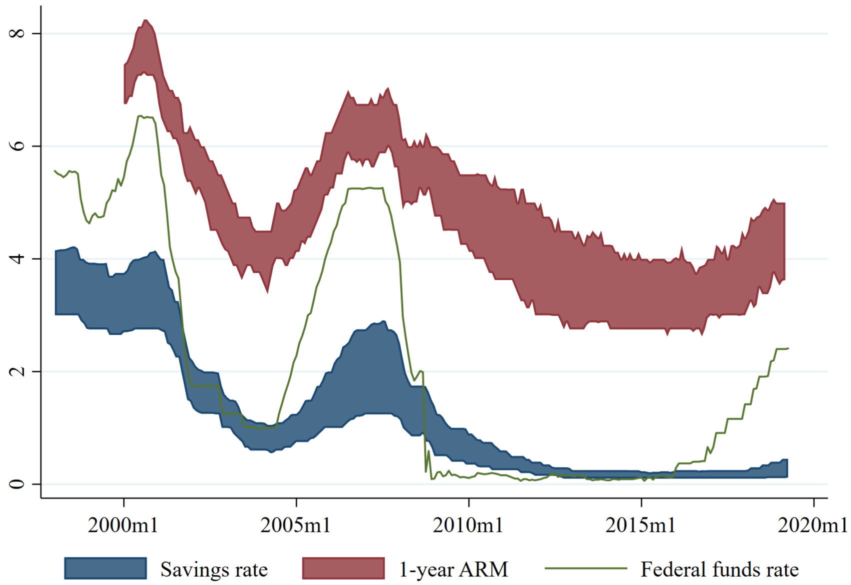
5.1 Rate Dispersion and Cyclical Spreads

Figure 3 presents the interquartile range (IQR) of the deposit and loan rates across all surveyed branches, along with the federal funds rate. Appendix A.4 offers similar evidence for a broader set of loan and deposit rates.[22]

[21] Drechsler et al. (2017) analyze deposit rates across locations but not across banks within a location.

[22] The focus on short-term rates abstracts from term premium effects. The 30-year fixed rate's cross-sectional dispersion is relatively small. Banks typically do not keep these loans on their balance sheets, selling or securitizing them. The ARM share was above 50% before 2007, then declined. Source: CoreLogic.

Figure 3: Deposit and loan rate IQR across bank branches



Notes: The shaded areas reflect the IQR of the 1-year adjustable loan rate and deposit rate for money market accounts with deposits of $25,000 for January 1998 to March 2019. The solid line represents the federal funds rate. Source: RateWatch, Federal Reserve Economic Data.

Fact 1: Dispersion within banks and locations. Bank loan and deposit rates are dispersed in the cross-section, both across locations within a bank institution and across institutions within a given location. The IQR measures total dispersion between 50 and 100 basis points in the cross-section but varied across time. LendingTree.com economists suggest consumers refinance their loans when the rate declines by about 50 basis points (see MarketWatch). Based on a mortgage of $175,000, the change yields an annual interest difference of $600 to $1,200. Both suggest that the observed cross-sectional dispersion is of economic significance and importance to households.

Telephone interviews with loan officers at large U.S. banks (e.g., Chase and PNC) suggest the institutions set prices strategically across locations depending on their local market share, and costs to originate loans vary across locations, explaining differences across branches of the same institution.[23] Table 2 shows the average loan and deposit rate dispersion (i.e., IQR) within locations and institutions. Focusing on loan rate dispersion in the upper part, within-location dispersion is higher than within-bank dispersion, at 1.03 versus 0.32, suggesting marginal costs play a more significant role than local concentration. The average deposit rate dispersion shown in the bottom part is smaller, at 0.57 and 0.21, for within-location and within-bank.[24]

[23] I conducted telephone interviews in the Summer of 2019 with loan officers at Chase and PNC.

[24] Rates tend to differ among commercial banks, savings & loan institutions (S&Ls), and credit unions, but

Table 2: Dispersion within-location and within-bank

	$m(\overline{IQR}_t^{loc})$	$m(\overline{IQR}_t^{bank})$
r_t^l	1.03	0.32
r_t^d	0.57	0.21

Notes: r_t^l reflects loan rate, r_t^d deposit rate, $m(\overline{IQR}_t^{loc})$ within-county average dispersion, and $m(\overline{IQR}_t^{bank})$ within-bank average dispersion. Source: RateWatch.

Fact 2: Countercyclical loan spreads. Spreads between branch-level loan rates and the federal funds rate tend to be high when the federal funds rate is low. The correlation between the average loan spread and federal funds rate is -0.84. The average spread is 3.57 for low federal funds rates and 1.8 for high federal funds rates, as shown in the left column of Table 3. Higher marginal costs and markups during low rate periods drive the differences across states. Section 3 and Section 6.2.3 explain why capital constraints are tighter during low federal funds rate periods. In contrast, the deposit spread between the federal funds rate and branch-level deposit rates is high when the federal funds rate is low. The correlation is 0.91. Similarly, the average deposit spread is 0.07 for low federal funds rates and 2.16 for high rates, implying that banks apply larger markdowns when interest rates are high.

Table 3: Spreads and dispersion for low and high federal funds rates

	$\rho(\bar{s}_t, r_t^f)$	$m(\bar{s}_t \vert r_t^f < 2)$	$m(\bar{s}_t \vert r_t^f \geq 2)$	$\rho(r_t^f, \overline{IQR}_t)$	$m(\overline{IQR}_t \vert r_t^f < 2)$	$m(\overline{IQR}_t \vert r_t^f \geq 2)$
r_t^l	-0.84	3.57	1.8	-0.57	1.33	1.06
r_t^d	0.91	0.07	2.16	0.88	0.36	1.11

Notes: ρ reflects the correlation coefficient of spreads, s_t, and the federal funds rate, r_t^f; m, is the conditional mean of loan rate, r_t^l, and deposit rate, r_t^d, IQRs during low, $(r_t^f < 2)$, and high, $(r_t^f \geq 2)$, federal funds rate periods. Source: RateWatch, Federal Reserve Economic Data.

Fact 3: Countercyclical rate dispersion. Loan and deposit rate dispersion varies with the federal funds rate. It moves in the same direction as the loan spread, indicating high loan rate dispersion for low federal funds rates and high deposit rate dispersion for high federal funds rates. The correlation between loan rate dispersion and the federal funds rate is -0.57, and 0.88 for deposit rate dispersion, as shown in the right column of Table 3. Similarly, loan rate dispersion is 27 basis points higher for low rates, while deposit rate dispersion is

adequate balance sheet data is not available for analysis beyond commercial banks and S&Ls.

75 points higher for high rates. The negative correlation between loan rate dispersion and the federal funds rate suggests that banks' marginal costs are more heterogeneous during low versus high rate periods, as capital requirements tighten.

Fact 4: Asymmetric adjustment. Pass-through asymmetry emerges between periods of monetary easing and tightening. While loan rates tend to adjust upwards quickly, they are downwards sticky, as indicated by slope differences observed between 2006, a period of monetary tightening, and 2008, a period of easing. Section 5.2 quantifies this relationship, and Section 6.2.3 explains the underlying mechanism.

5.2 Monetary Policy Pass-Through in Cross-Section and Time Series

This section examines pass-through dynamics using local projection methods (Jordà, 2005), as they provide a flexible framework and allow for heterogeneity and asymmetry. The analysis focuses on the speed and extent of monetary policy pass-through, i.e., how fast and completely banks pass changes in costs to consumers. To capture the relative importance of local bank concentration and capitalization, the variables are interacted with the shock.

The baseline model estimates the pass-through of monetary policy shocks to loan rates at each horizon, $h \in [0, H]$, by regressing branch i's retail rate adjustment, $r^l_{t+h,i,c} - r^l_{t-1,i,c}$, on the monetary policy shock, s_t, interacted with the variable of interest, $X_{t-1,i,c}$:[25]

$$r^l_{t+h,i,c} - r^l_{t-1,i,c} = \alpha^h_i + \beta^h s_t + \underbrace{\gamma^h \; s_t \; \times \; X_{t-1,i,c}}_{\substack{\text{local HHI or} \\ \text{bank capitalization}}} + \theta^h X_{t-1,i,c} + \eta^h Z_{t,c} + \epsilon_{t+h,i,c} \tag{6}$$

where $r^l_{t+h,i,c} - r^l_{t-1,i,c}$ reflects the loan rate change between $t+h$ and $t-1$. The regression is estimated for each horizon h and includes branch fixed effects, α^h_i, controls for national and local economic conditions, $Z_{t,c}$.[26] To address endogeneity concerns, I use the lagged values of the interaction variables.[27] The main coefficient of interest in equation (6) is γ^h, the local HHI or capitalization's marginal effect on pass-through. β^h serves as reference point to indicate average pass-through.[28] To interpret bank concentration and capitalization's effects, the impulse responses are presented for high and low states, defined as two

[25] The regression includes the interaction terms jointly. The results hold including the variables individually.

[26] The set of controls includes two lags of the county-level and national unemployment rate, real GDP growth, CPI inflation, county-level median debt-to-income ratio, dependent variable and monetary shock, and a dummy for the zero lower bound-period.

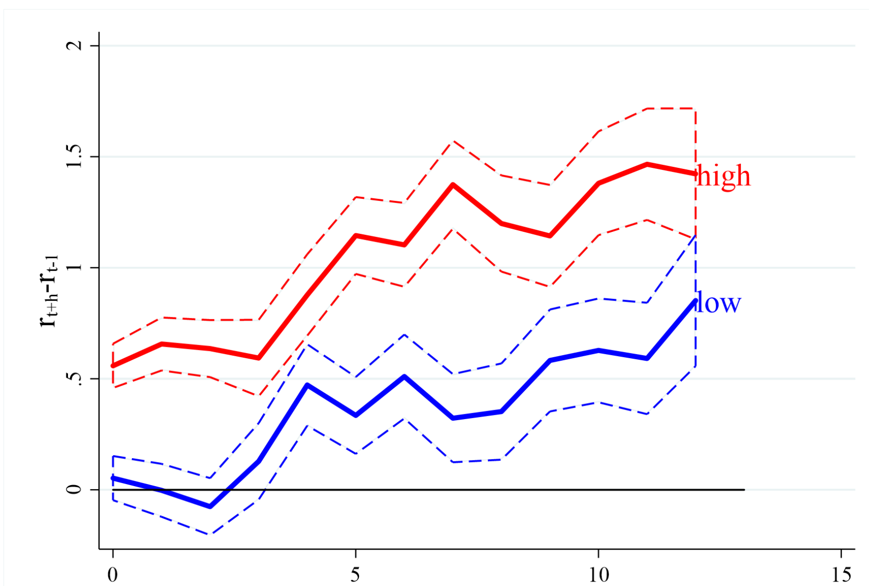
[27] The focus lies on cross-sectional differences, not on time differences. To control for time trends in the underlying bank capital ratio variable, the analysis uses deviation from the period average.

[28] Adding time dummy variables yields qualitatively similar results but provides no benchmark.

standard deviations above or below the mean of characteristic, $X_{t,i,c}$.[29] The representation simplifies interpretation but maintains a continuous interaction term. The monetary shock is scaled to increase the federal funds rate by 1 percentage point on impact.[30]

Local bank concentration. Figure 4 presents impulse response functions for loan rates to a monetary shock at both a high and low bank concentration level. High-concentration branches adjust loan rates more in response to the shock than low-concentration branches by about 50 basis points on impact and increasing over ensuing months. In the low-concentration region, pass-through is incomplete, i.e., less than one after 12 months. The findings are consistent with the predictions from the heterogeneous pass-through model in Section 3. Banks operating in high-concentration markets serve customers with relatively low demand elasticity and exhibit high market power, leading to higher loan spreads and pass-through. The divergence of loan rates across branches in response to a monetary shock also explains a widening dispersion during policy changes in Figure 3.

Figure 4: Impulse responses of loan rates by local bank concentration



Notes: Impulse response functions of 1-year hybrid ARM rates to a monetary policy shock at both high and low local bank concentrations, calculated as $\beta^h + \gamma^h\ (m^{HHI} \pm 2sd^{HHI})$. Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

Bank capitalization. Previous research (e.g., Kashyap and Stein, 2000) finds that low capitalization banks or banks with relatively illiquid balance sheets respond more to monetary policy. Similarly, the simple model in Section 3 predicts that banks with a

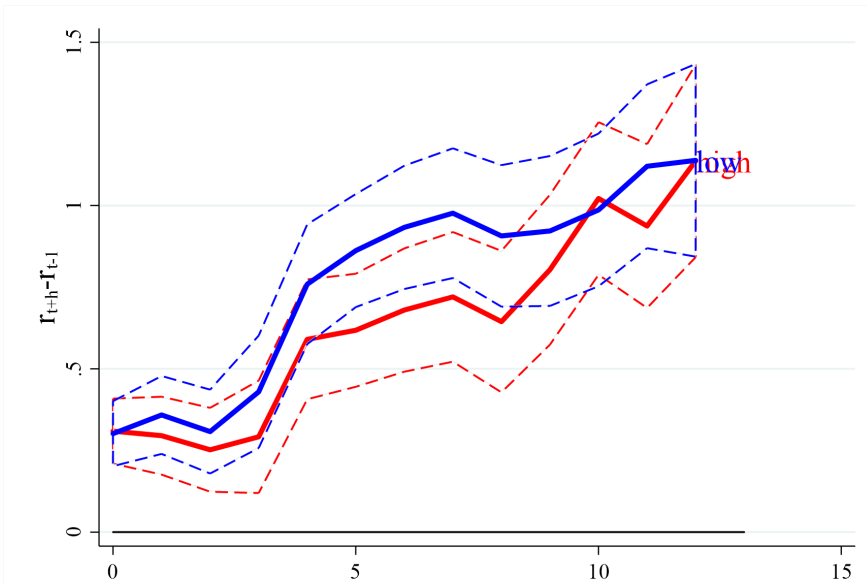
[29]The high (low) pass-through is calculated as $\beta^h + \gamma^h\ (m^X \pm 2sd^X)$.

[30]I regress the federal funds rate change on the shock and use the coefficient as a scaling parameter.

relatively low bank capital ratio will adjust loan rates more to changes in funding costs and benefit less from capital constraint easing.

Figure 5 shows loan rate impulse response functions to a monetary shock for low and high bank capital ratios. The figure demonstrates greater pass-through for banks with a low, versus high, capital ratio, in line with the simple model. However, bank capitalization seems to play a lesser role than concentration; there is a smaller difference in impulse responses, and the confidence intervals overlap. The temporary divergence of loan rates across banks in response to a monetary shock also explains a widening dispersion during monetary policy changes in Figure 3.

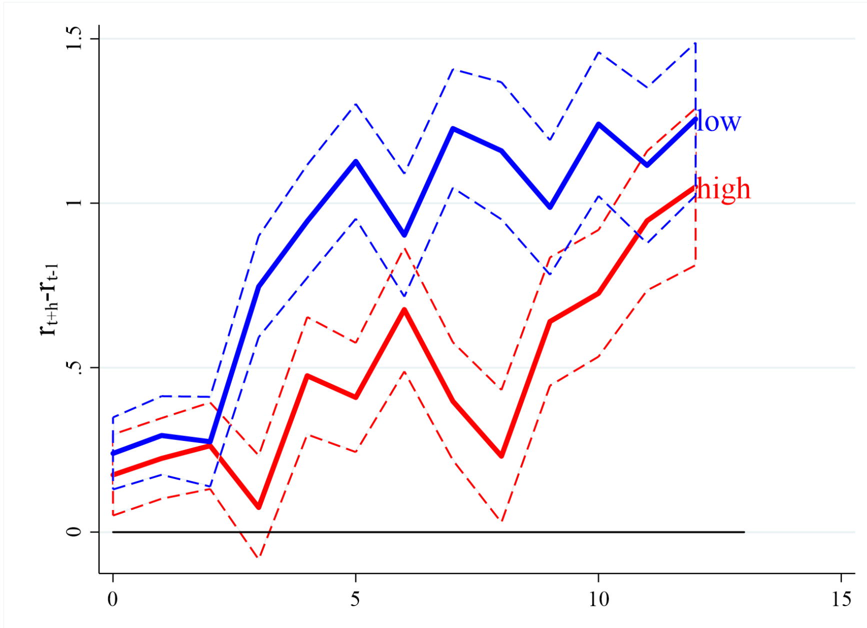
Figure 5: Impulse responses of loan rates by bank capital ratio



Notes: Impulse response functions of 1-year hybrid ARM rates to a monetary policy shock at both high and low capitalization. The functions are calculated as $\beta^h + \gamma^h\ (m^{\%} \pm 2sd^{\%})$. Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

System-wide capitalization. In addition to heterogeneity between bank institutions, capitalization varies over time with the credit cycle. In times of systemic stress, such as during the Great Financial Crisis of 2009, lower aggregate capitalization could additionally impact pass-through. What is the role of capitalization as a function of system-wide capital? To examine this question, consider the interaction of the aggregate capital ratio with the monetary policy shock. Figure 6 shows the loan rate impulse responses to a monetary policy shock given a low and high system-wide capital. In times of financial stress and low system-wide capital, monetary pass-through is higher. Like the capitalization of individual banks, system-wide capitalization affects pass-through, leading to a faster adjustment.

Figure 6: Impulse responses of loan rates by system-wide capitalization



Notes: Impulse response functions of 1-year hybrid ARM rates to a monetary policy shock at both high and low system-wide capitalization. The functions are calculated as $\beta^h + \gamma^h\ (m^{\%} \pm 2sd^{\%})$. Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

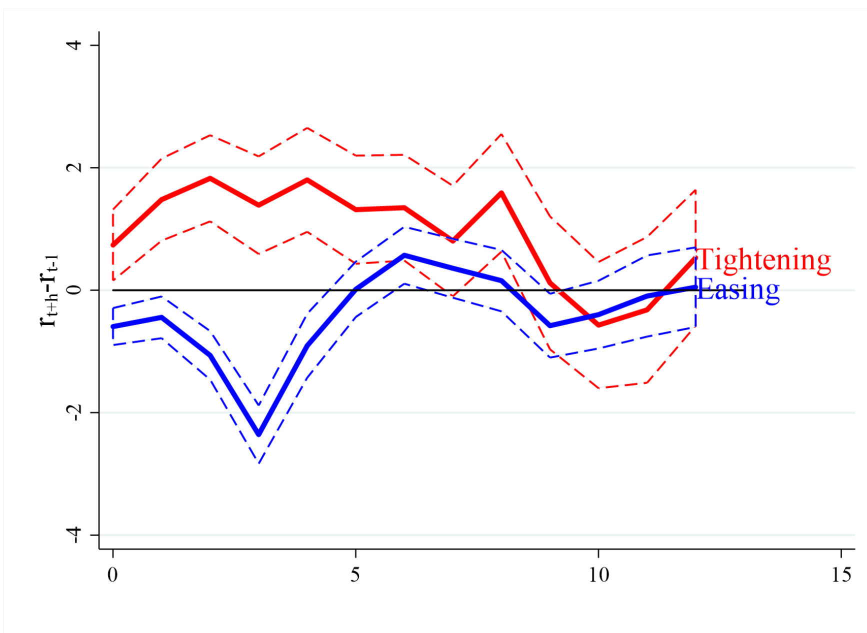
Increasing aggregate pass-through over time. The rise in bank concentration at the local level and the reduction in aggregate bank capitalization point to an amplification of monetary policy transmission in recent decades. Consistent with the cross-sectional projections, sub-sample analyses based on aggregated data suggest that monetary policy pass-through to mortgage rates and real estate loans has increased over time. Figure A.14 in Appendix A.7 presents impulse responses of the aggregate mortgage rate and real estate loans to monetary policy surprises for the early and late periods documenting that aggregate pass-through to mortgage rates is higher in the second period than in the first. Similarly, real estate loans from commercial banks contract by more in the second period, in line with the predictions from the cross-section.

Monetary tightening vs. easing. Building on the evidence for a greater pass-through during periods of monetary tightening versus easing in Figure 3, I assess the state-dependency of monetary policy pass-through. I interact the monetary policy shock, s_t, with an indicator for periods with expected monetary tightening, $\mathbb{I}(\mathbb{E}_{t-1}\Delta r_t^f > 0)$, and for periods of expected monetary easing, $\mathbb{I}(\mathbb{E}_{t-1}\Delta r_t^f < 0)$. I define the expected change in the federal funds rate, $\mathbb{E}_{t-1}\Delta r_t^f$, as the actual change minus the realized monetary shock:[31]

$$r^l_{t+h,i,c} - r^l_{t-1,i,c} = \alpha_i^h + \beta^h s_t + \underbrace{\mathbb{I}\left(\mathbb{E}_{t-1}\Delta r_t^f > 0\right)}_{\text{tightening}}\left(\alpha_i^{h,+} + \beta^{h,+} s_t\right) \tag{7}$$
$$+ \underbrace{\mathbb{I}\left(\mathbb{E}_{t-1}\Delta r_t^f < 0\right)}_{\text{easing}}\left(\alpha_i^{h,-} + \beta^{h,-} s_t\right) + \eta^h Z_{c,t} + \epsilon_{t+h,i,c}$$

Figure 7 confirms that pass-through is greater during monetary tightening than easing, which shows a negative response. Hence, the loan rate increases with negative monetary shock during easing periods, implying a negative pass-through. Appendix A.8 provides an extension with double interaction terms and shows that the bank concentration and capitalization results hold in both sub-periods.

Figure 7: Impulse responses of loan rates by monetary easing vs. tightening



Notes: Impulse response functions of 1-year hybrid ARM rates to a monetary policy shock during expected monetary tightening and easing periods. Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

[31]Using raw changes in the federal funds rate yields similar results.

6 Quantitative Model

This section introduces a dynamic stochastic general equilibrium model to quantify the relative importance of market power, capital requirements, and adjustment costs for monetary policy pass-through. Using counterfactual analyses, I then assess the impact of rising bank concentration on monetary policy pass-through and monetary transmission to the real economy. The model builds on Gerali et al. (2010) and features standard New Keynesian building blocks. The model assumes segmented financial markets, where patient households provide deposits to the banking sector and impatient households and entrepreneurs demand credit for investment in housing and capital. A monetary authority sets the policy rate via a Taylor rule. As in the simple model in Section 3, banks operate in an environment with monopolistic competition in deposit and loan markets and face a capital requirement. In addition, banks are subject to quantity adjustment costs on loans and deposits. The remaining building blocks follow Gerali et al. (2010). See Appendix B.1 and B.2 for model details beyond the banking sector and Appendix B.3 for the calibration.

6.1 The Banking Sector

Following Gerali et al. (2010), the banking sector is divided into three parts: a representative wholesale management unit (comparable to bank headquarters), a continuum of retail deposit branches, and retail loan branches operated under monopolistic competition.

6.1.1 Wholesale Unit

The representative wholesale unit manages funds between retail deposit and loan branches and is subject to a bank capital requirement. The wholesale unit's total bank lending, B_t, is composed of retail branches financing loans to households, b_t^{bH}, and entrepreneurs, b_t^{bE}, with $B_t = b_t^{bH} + b_t^{bE}$. Its liabilities are composed of funds from deposit branches, d_t^p, and bank capital, K_t^b. The wholesale unit retains previous period's profit to cover incidental management costs. As a result, bank capital, K_t^b, evolves as:

$$\pi_t K_t^b = \left(1 - \delta^b\right) K_{t-1}^b + \Pi_{t-1}^b, \tag{8}$$

where Π_{t-1}^b reflects retained profits, δ^b the required resources for managing bank capital, and π_t the inflation rate. Any deviation from the required bank capital is modeled with a quadratic cost function, $\mathbb{A}_{KB}\left(\frac{K_t^b}{B_t}\right) = \frac{\kappa_{KB}}{2}\left(\frac{K_t^b}{B_t} - \nu^b\right)^2$, governed by cost parameter κ_{KB}, instead of explicitly modeling the capital constraint, which avoids non-linearities while otherwise similar (Brunnermeier and Koby, 2018; Gerali et al., 2010).

The wholesale unit generates income from providing wholesale funding to its retail loan branches, B_t, at the wholesale funding rate, R_t^b, minus expenses paid to its retail deposit branches, d_t^p, at the wholesale lending rate, R_t^d. The wholesale lending rate, R_t^d, equals the central bank policy rate, r_t^f, in equilibrium. The wholesale unit discounts future profits with the stochastic discount factor of the patient household, $\Lambda_{0,t}^P$, and maximizes:

$$\max_{B_t,d_t^p} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[R_t^b B_t - R_t^d d_t^p - \mathbb{A}_{KB} \left(\frac{K_t^b}{B_t} \right) K_t^b \right], \tag{9}$$

subject to the wholesale unit's balance sheet constraint:

$$B_t = d_t^p + K_t^b. \tag{10}$$

Solving the wholesale unit's maximization problem and rewriting the first-order condition yields the wholesale funding rate as a function of bank capital ratio, ν^b, and policy rate, r_t^f:

$$R_t^b = r_t^f - \kappa_{KB} \left(\frac{K_t^b}{B_t} - \nu^b \right) \left(\frac{K_t^b}{B_t} \right)^2. \tag{11}$$

Equation (11) indicates the loan rate depends inversely on the bank capitalization outside the steady state, as in the simple model of heterogeneous pass-through in Section 3. The cost term in parentheses becomes negative when the policy rate decreases, as expanding bank lending increases B_t by more than K_t^b. The more so, the higher the cost parameter, κ_{KB}, and steady-state bank capital ratio, ν^b. Banks target the steady-state bank capital ratio; hence, the term in parentheses becomes zero in the steady state.

6.1.2 Retail Deposit Branches

Retail deposit branches collect deposits from patient households and store these at the wholesale unit at the wholesale lending rate, R_t^d. The deposit branches earn a positive spread on the deposit rate due to monopolistic deposit market competition. Deposit branches incur adjustment costs from changing deposits, as attracting new customers requires additional processing and advertising. Flannery (1982) regards deposits as "quasi-fixed" inputs, which may also explain why deposit rates exceeded the federal funds rate for some time periods in Figure 3. The adjustment costs, $\mathbb{A}_D$, are expressed as deviations from the steady-state deposit level, d_{ss}^p, and take the form: $\frac{\kappa_d}{2} \left(\frac{d^p(r_t^d)}{d^p(r_{ss}^d)} - 1 \right)^2$, governed by cost parameter κ_d. Each deposit branch maximizes its discounted future profits as follows:[32]

$$\max_{r_t^d} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[R_t^d d^p(r_t^d) - r_t^d d^p(r_t^d) - \mathbb{A}_D \left(d^p(r_t^d) \right) \bar{r}_t^d \bar{d}_t^p \right], \tag{12}$$

subject to the local deposit supply function:

$$d^p(r_t^d) = \left(\frac{r_t^d}{\bar{r}_t^d} \right)^{-\epsilon^d} \bar{d}_t^p, \tag{13}$$

where $\bar{r}_t^d$ and $\bar{d}_t^p$ reflect the aggregate deposit rate and deposits. After imposing symmetry, ($d_t^p = \bar{d}_t^p$, $r_t^d = \bar{r}_t^d$), the deposit branch's optimality condition is:

$$-\epsilon^d \frac{R_t^d}{r_t^d} + \left(\epsilon^d - 1 \right) + \epsilon^d \kappa_d \left(\frac{d_t^p}{d_{ss}^p} - 1 \right) \frac{d_t^p}{d_{ss}^p} = 0 \tag{14}$$

The branch determines the deposit rate based on (i) deposit supply elasticity, ϵ^d, (ii) wholesale lending rate, R_t^d (which equals the policy rate, r_t^f), and (iii) deviation from the steady-state deposit level. Accordingly, cross-sectional heterogeneity may emerge in deposit rates due to differences in deposit supply elasticity ϵ^d, as shown in the simple model, and adjustment costs, κ_d, or the steady-state deposit level (i.e., branch size).

6.1.3 Retail Loan Branches

Retail loan branches of type l, with $l \in \{bH, bE\}$, finance loans to impatient households, b_t^{bH}, or entrepreneurs, b_t^{bE}, with funding from the wholesale unit at a wholesale funding rate, R_t^b. Similar to the retail deposit branches, retail loan branches earn a positive spread due to monopolistic loan market competition. Each loan branch incurs costs from adjusting lending, $\mathbb{A}_l$. Anecdotal evidence suggests banks struggle to increase lending during periods of low interest rates and high loan demand, implying higher adjustment costs during loan expansions. An altered linear exponential loss function is used here to generate asymmetry (Abbritti and Fahr, 2013; Fahr and Smets, 2010; Levieuge and Sahuc, 2021). Adjustment costs are defined in terms of deviations from the steady-state loan level, b_{ss}^l, and take the form: $\frac{\kappa_l}{2} \left(\frac{b_t^l}{b_{ss}^l} - 1 \right)^2 + \frac{1}{\psi_l^2} \left\{ \exp \left[\psi_l \left(\frac{b_t^l}{b_{ss}^l} - 1 \right) \right] - \psi_l \left(\frac{b_t^l}{b_{ss}^l} - 1 \right) - 1 \right\}$, where parameters κ_l and ψ_l govern convexity and asymmetry. $\psi_l > 0$ generates higher costs when lending is above the steady state, i.e. $\left(\frac{b_t^l}{b_{ss}^l} - 1 \right) > 0$. When ψ_l approaches 0, the function nests the symmetric case. Appendix B.5 describes the cost function's micro-foundation.

[32] Retail branches discount future profits with the patient household's stochastic discount factor $\Lambda_{0,t}^P$.

Each loan branch maximizes its discounted future profits as follows:[33]

$$\max_{r_t^l} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[r_t^l b_t^l(r_t^l) - R_t^b b_t^l(r_t^l) - \mathbb{A}_l \left(b_t^l(r_t^l) \right) \overline{r}_t^l \overline{b}_t^l(r_t^l) \right] \tag{15}$$

subject to the local loan demand function:

$$b_t^l(r_t^l) = \left(\frac{r_t^l}{\overline{r}_t^l} \right)^{-\epsilon^l} \overline{b}_t^l \;\; \forall \;\; l \in \;\; \{bH, bE\} \tag{16}$$

where $\overline{r}_t^l$ and $\overline{b}_t^l$ reflect aggregate loan rate and loans. After imposing symmetry, the loan branch's optimality condition is $\forall \;\; l \in \;\; \{bH, bE\}$:

$$-\left(\epsilon^l - 1\right) + \epsilon^l \frac{R_t^b}{r_t^l} + \epsilon^l \kappa_l \left(\frac{b_t^l}{b_{ss}^l} - 1 \right) \frac{b_t^l}{b_{ss}^l} + \frac{\epsilon^l}{\psi_l} \left\{ \exp\left[\psi_l \left(\frac{b_t^l}{b_{ss}^l} - 1 \right) \right] - 1 \right\} \frac{b_t^l}{b_{ss}^l} = 0 \tag{17}$$

The loan rate decision is determined by: (i) loan demand elasticity, ϵ^l, (ii) wholesale funding rate, R_t^b, and (iii) loan portfolio changes. The exponential function collapses to zero when the loan volume declines, generating state-dependent effects conditional on policy rate easing or tightening. The loan rate setting equation suggests that heterogeneity in monetary policy pass-through to retail rates can be explained by differences in market power, ϵ^l, adjustment costs, κ_l and ψ_l, steady-state loans volumes (ie., branch size), and bank capital constraints, ν^b and κ_{KB}.

6.2 Comparative Statics

To determine how monetary policy pass-through changes if banks (i) have more market power, (ii) must fulfill a greater bank capital requirement, or (iii) incur higher adjustment costs, I compare impulse response functions to a monetary shock across parameterizations, similar to the previous empirical analysis. The approach also explains the mechanics of the *market power channel* and the *capital allocation channel*.

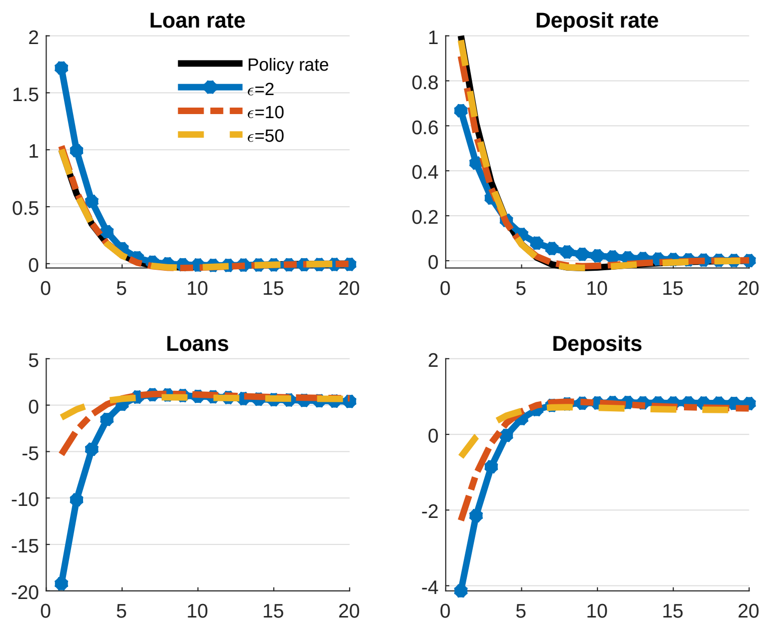
6.2.1 Market Power

I examine the impulse response functions of loan rate, deposit rate, aggregate household loans, and aggregate deposits to a monetary policy shock varying the elasticities of deposit supply, ϵ^d, and loan demand, ϵ^l, while holding all other parameters constant. The monetary shock is scaled to increase the policy rate on impact by 1 percentage point, as in the

[33] Loan branches discount future profit with the patient household's stochastic discount factor $\Lambda_{0,t}^P$.

empirical section. Figure 8 shows that the lower ϵ^l and ϵ^d (in absolute terms) in conjunction with higher market power, the higher the pass-through to loan rate and the lower the pass-through to deposit rate.

Figure 8: Impulse responses to a monetary tightening varying ϵ^d and ϵ^l



Notes: Impulse responses to a monetary shock varying deposit supply and loan demand elasticity (ϵ^d, ϵ^l).

In response to a policy rate increase by 1 percentage point, the loan rate increases by almost a factor of 1.75 in the high market power case shown in the upper left panel, broadly in line with the empirical results. The deposit rate increases by about 65 basis points, or a factor of 0.65, in the high market power case in the figure's upper right panel. Similarly, Drechsler et al. (2017) find that branches operating in high-concentration markets increase deposit rates by less. The figure's bottom panels present results for household loans and deposits. As ϵ declines, both respond by more amplifying the credit cycle. Concretely, household loans decline in the high market power case by 20% compared to less than 2% in the low market power case.

Consider the linearized loan and deposit rate-setting equations:

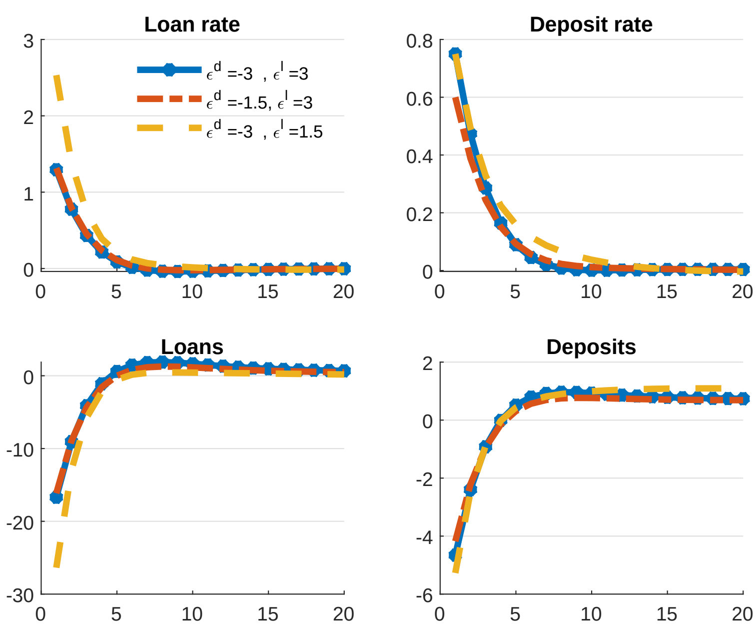
$$\widehat{r}_t^l = \underbrace{\frac{\epsilon^l}{(\epsilon^l - 1)}}_{\text{markup}} \widehat{R}_t^b + \underbrace{\frac{\epsilon^l}{(\epsilon^l - 1)}}_{\text{markup}} \kappa_l r^l \widetilde{b}_t^l, \tag{18}$$

$$\widehat{r}_t^d = \underbrace{\frac{\epsilon^d}{(\epsilon^d - 1)}}_{\text{markdown}} \widehat{r}_t^f + \underbrace{\frac{\epsilon^d}{(\epsilon^d - 1)}}_{\text{markdown}} \kappa_d r^d \widetilde{d}_t^p, \tag{19}$$

where $\widehat{r}_t^l$, $\widehat{r}_t^d$, $\widehat{r}_t^f$ and $\widehat{R}_t^b$ are expressed as absolute deviations and $\widetilde{b}_t^l$ and $\widetilde{d}_t^p$ as percentage deviations from their steady-state values. Equations (18) and (19) indicate loan and deposit rates increase proportionally to loan markup and deposit markdown in absence of adjustment costs (i.e., setting κ_l and κ_d to zero). In the presence of adjustment costs, the effect of market power is attenuated; see Section 6.2.3 for more details.

After discussing the comparative statics for simultaneously changing the elasticities of loan demand and deposit supply in Figure 8, I examine the impact of changing only one to gain insight into which is more important. Figure 9 presents the comparative statics holding either the elasticity of deposit supply or loan demand constant while varying the other. While higher loan market power increases loan rate pass-through and amplifies the credit cycle, higher deposit market power minimally alters loan rate pass-through and the credit cycle. The effect suggests that considering deposit market power alone as Drechsler et al. (2017) is insufficient for explaining lending movements due to higher market concentration.

Figure 9: Impulse responses to a monetary tightening varying ϵ^d and ϵ^l

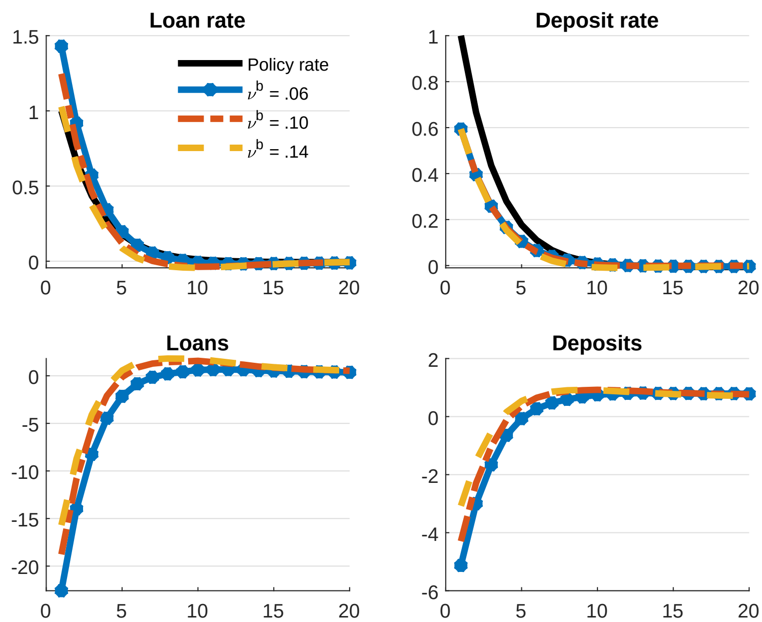


Notes: Impulse responses to a monetary shock varying deposit supply and loan demand elasticity (ϵ^d, ϵ^l).

6.2.2 Bank Capital Ratio

To analyze bank capital's role in pass-through and examine the *capital allocation channel*, I vary the bank capital ratio, ν^b, that the bank holds in the steady state. Figure 10 displays impulse response functions of loan rate, deposit rate, aggregate household loans, and aggregate deposits to a monetary policy shock across different parameterizations of bank capital ratio, ν^b, while holding all other parameters constant. A low bank capital ratio increases pass-through to loan rates, in line with the empirical results. Similarly, aggregate bank lending responds more when banks hold a lower bank capital ratio than when the ratio is high, implying that the credit cycle is more affected. In contrast to loan rates, deposit rate pass-through is not affected by bank capital ratio changes.

Figure 10: Impulse responses to a monetary tightening varying ν^b



Notes: Impulse response functions to a monetary shock varying bank capital ratio ν^b.

Consider the linearized wholesale funding rate in equation (11), which is proportional to the loan rate:[34]

$$\widehat{R}_t^b = \widehat{r}_t^f - \kappa_{KB} \left(\nu^b\right)^3 \left(\widetilde{K}_t^b - \widetilde{B}_t\right), \tag{20}$$

where $\widetilde{K}_t^b$ and $\widetilde{B}_t$ are expressed as percentage deviations from their steady-state values. Equation (20) shows the wholesale funding rate, $\widehat{R}_t^b$, as a function of bank capital ratio, ν^b, and the gap between bank capital and loans, $\left(\widetilde{K}_t^b - \widetilde{B}_t\right)$. The gap becomes negative in response to a negative monetary policy shock because lending, $\widetilde{B}_t$, expands more than bank capital, $\widetilde{K}_t^b$ in response to a shock.[35] Hence, the wholesale funding rate, $\widehat{R}_t^b$, declines less than policy rate, $\widehat{r}_t^f$, the more so the lower the bank capital ratio. Equation (11) is similar to equation (1) from the simple model, as both depend inversely on the bank capital requirement.

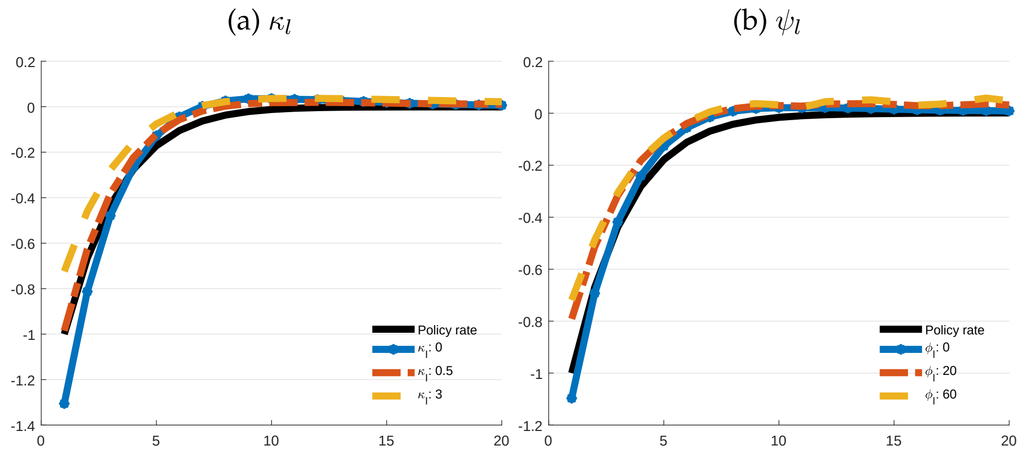
[34]The wholesale funding rate equals the loan rate times the inverse markup: $\widehat{R}_t^b \approx \widehat{r}_t^l \dfrac{(\epsilon^l - 1)}{\epsilon^l}$.

[35]Current bank capital equals previous period's capital and profits minus management costs (Equation (8)).

6.2.3 Adjustment Costs

Why do retail rates slowly and incompletely adjust to monetary shocks? I consider the role of adjustment costs via comparative statics for two parameters, κ_l, and ϕ_l, which govern the cost's convexity and symmetry in Equation (17). Because loan adjustment costs do not affect deposit rates, I focus solely on loan rates. Panels (a) and (b) in Figure 11 present loan rate impulse responses to a negative monetary shock varying κ_l and ϕ_l along with policy rate's impulse response function as a comparison. Loan rate pass-through declines with increasing adjustment costs, leading to an incomplete pass-through.

Figure 11: Loan rate impulse responses to a monetary easing varying κ_l and ψ_l



Notes: Impulse response functions to a negative monetary shock varying the shape of the adjustment cost function. κ_l changes the degree of convexity and ϕ_l the symmetry.

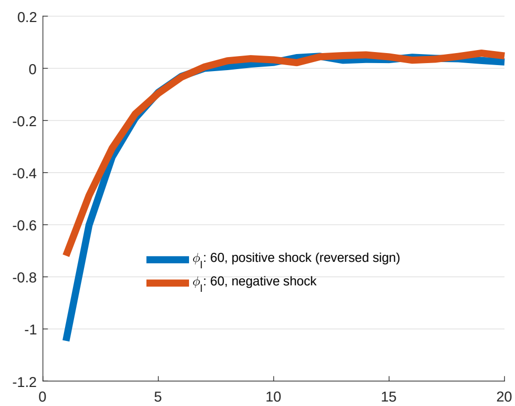
To understand the adjustment cost mechanism, consider the cross partial derivative of the linearized loan rate equation (18) to wholesale funding rate, $\widehat{R}_t^b$, and cost, κ_l:

$$\frac{\partial \widehat{r}_t^l}{\partial \widehat{R}_t^b \partial \kappa_l} = \frac{\epsilon^l}{\epsilon^l - 1} r^l \frac{\partial \widehat{b}_t^l}{\partial \widehat{R}_t^b} < 0. \tag{21}$$

Equation (21) indicates pass-through declines with rising adjustment costs, κ_l, as lending falls with a rising rates, i.e., $\frac{\partial \widehat{b}_t}{\partial \widehat{R}_t^b}$ is negative. Market power has an amplifying role, implying that adjustment costs dampen and counteract the channel's direct impact.

Figure 12 compares the impulse responses to a positive and negative shock. While the loan rate declines by about 100 basis points in response to a positive shock, it only declines 75 basis points with a similar-size negative shock. The asymmetry is due to bank's costs for expanding the loan portfolio, creating a sluggish downward rate adjustment. The results on quantity adjustment costs are qualitatively similar to findings on price adjustment costs (Levieuge and Sahuc, 2021). However, anecdotal evidence favors quantity over price adjustment costs, as banks effectively incur higher charges of expanding lending (e.g., additional overhead, screening costs). For more details, see Appendix B.5.

Figure 12: Loan rate impulse responses to monetary tightening versus easing



Notes: Impulse response functions to a negative and positive (reversed-sign) monetary shock and asymmetric adjustment costs.

An alternative explanation for asymmetric monetary policy pass-through builds on the intuition that banks facing a minimum capital ratio incur greater costs when undershooting than overshooting. Hence, introducing asymmetric bank capital adjustment costs at the headquarters level leads to a comparable impact: Loan rates decline less in response to monetary easing. Figure B.3 in Appendix B.5 shows the equivalence of both approaches. In reality, both channels likely function simultaneously and reinforce each other.

6.3 Quantitative Assessment of Rise in Bank Concentration

This section quantifies the implications of rising bank concentration for monetary policy pass-through using counterfactual analyses. In this, I distinguish between the *market power channel*, changes in the underlying *market environment*, and the *capital allocation channel*, shifts in the composition of the *banking sector*. I expand the model to include heterogeneous bank branches operating in spatially segmented markets and belonging to heterogeneous bank headquarters. Specifically, bank branches operate in local markets with varying market power, and their bank headquarters hold size-dependent bank capital ratios. To capture bank heterogeneity in a tractable framework, assume two types along each dimension: regional banks, and giant banks, denoted by the superscripts r and g, paired with a continuum of branches in low- and high-concentration markets, denoted l and h. The approach yields four types of bank branches: (i) Regional banks in low-concentration markets,

(ii) regional banks in high-concentration markets, (iii) giant banks in low-concentration markets, and (iv) giant banks in high-concentration markets. Correspondingly, there is a share of branches operating in high-concentration markets, α^m, and giant banks, α^b. Table 4 shows the derived bank branch-specific loan rates depending on local concentration, $\epsilon^m \; \forall \; m \; \in \; \{l, h\}$, and headquarters-specific marginal costs, $R_t^j \; \forall \; j \; \in \; \{r, g\}$.

Table 4: Heterogeneous bank headquarters and markets

		Bank types		
		Regional	Giant	Share
Local market concentration	Low	$r_t^{l,r} = \frac{\epsilon^l}{\epsilon^l - 1} R_t^r$	$r_t^{l,g} = \frac{\epsilon^l}{\epsilon^l - 1} R_t^g$	α^m
	High	$r_t^{h,r} = \frac{\epsilon^h}{\epsilon^h - 1} R_t^r$	$r_t^{h,g} = \frac{\epsilon^h}{\epsilon^h - 1} R_t^g$	$(1 - \alpha^m)$
	Share	α^b	$\left(1 - \alpha^b\right)$	

Notes: Branch-specific loan rates depend on local concentration, $\epsilon^m \; \forall \; m \; \in \; \{l, h\}$, and headquarters-specific marginal costs, $R_t^j \; \forall \; j \; \in \; \{r, g\}$. $(1 - \alpha^m)$ refers to high-concentration; and $\left(1 - \alpha^b\right)$ giant banks' share.

In the counterfactual analyses, I contrast monetary policy pass-through in a calibrated banking sector for 1994 and 2019, representing relatively low and high bank concentration environments. Specifically, I consider changes along the *extensive* margin, i.e., the share of high-concentration markets, $(1 - \alpha^m)$, and giant banks, $\left(1 - \alpha^b\right)$, in line with U.S. trends presented in Appendix B.4. The first scenario increases the share of high-concentration markets, $(1 - \alpha^m)$, matching shifts in relative market size for high-concentration counties. The second scenario increases the market share of giant banks, $\left(1 - \alpha^b\right)$, matching shifts in bank headquarters distribution. Finally, I explore the combined effects of market structure and bank composition changes. Further, I account for trends in markups and bank capital ratios over time, corresponding to the *intensive* margin.

Table 5: Heterogeneous banks model calibration

	Parameter	α^m	α^b	ϵ^d	$\epsilon^{bH/E}$	ν^b
1994	Bank/Branch I	0.7	0.9	-2.60	2.51	0.09
	Bank/Branch II	0.3	0.1	-1.03	2.05	0.06
2019	Bank/Branch I	0.4	0.4	-0.99	1.68	0.12
	Bank/Branch II	0.6	0.6	-0.32	1.46	0.09

Notes: The row Branch/Bank I (Bank/Branch II) presents the calibration of ϵ^d, ϵ^{bH}, ϵ^{bE} and ν^b for the low-concentration market and regional bank (high-concentration market and giant bank) by period, 1994 and 2019. α^m and α^b reflect the share of low-concentration markets and regional banks, respectively.

Table 5 presents the calibration details for the different scenarios. I calibrate low-concentration markets' share, α^m, regional banks' share, α^b, deposit supply elasticity, ϵ^d, elasticities of loan demand from households, ϵ^{bH}, and entrepreneurs, ϵ^{bE}, and bank capital requirement, ν^b, separately for low- and high-concentration markets and regional and giant banks, as well as two periods, 1994 and 2019.[36] First, I calibrate low-concentration markets' share, α^m, and deposit supply elasticity, $\epsilon^{d,c}$, and loan demand elasticity, $\epsilon^{j,c} \; \forall \; j \; \in \{bH, bE\}$ for market $c \in \{l, h\}$. α^m is derived from the county-level HHI distribution across time. $\epsilon^{j,c} \; \forall \; j \quad \in \{d, bH, bE\}$ is inferred from bank-level interest income and expense data and calibrated to the average cross-sectional, asset-weighted markups/markdowns and dispersion.[37] Second, giant banks are defined as those above \$100 billion in assets (in \$2018). I calculate giant banks' share, $(1 - \alpha^b)$, and the annual weighted group means of the bank capital ratio, ν^b, separately for giant and regional banks, defined as those with assets below \$100 billion.

6.3.1 Heterogeneous Bank Branches: Rising High-Concentration Markets

How do market structure changes affect aggregate retail rates and monetary policy pass-through, particularly an increase in the share of high-concentration markets? The empirical section establishes that banks operate in several local markets and choose location-specific deposit and loan rates. I consider two spatially segmented branch types with differing loan demand and deposit supply elasticities to capture heterogeneity within a bank across markets. Appendix B.6 describes the modifications and includes analytical proofs.

Propositions 1 and 2 suggest the heterogeneous retail rates simplify to a sufficient statistic depending only on exogenous parameters, α^m, ϵ^l, and ϵ^h, and offer pass-through insight.

Proposition 1. *The aggregate markup provides a sufficient statistic summarizing the degree of heterogeneity between branches and market shares and informs about monetary policy pass-through.*

Proposition 2. *The aggregate loan rate pass-through decreases in low-concentration share, α^m, and the sensitivity depends on the degree of heterogeneity, the difference between ϵ^l and ϵ^h. Further, it decreases in high-concentration markets' elasticity, ϵ^h. For the deposit side, the opposite holds.*

[36] The 1994 and 2019 calibration rely on bank data for the periods 2000-2008 and 2009-2019.

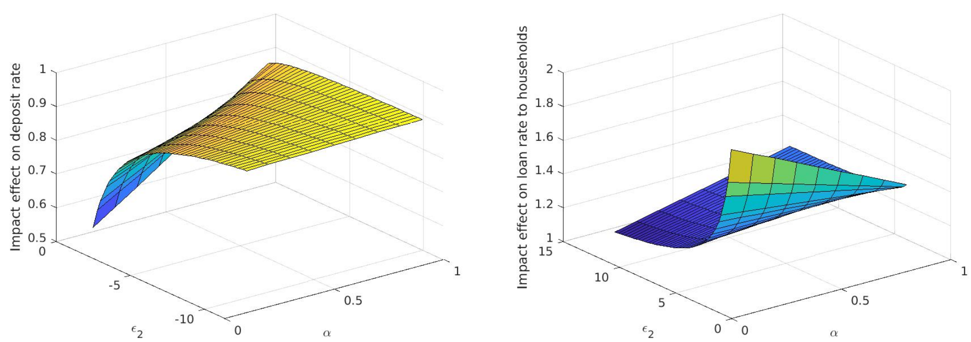
[37] The markup/markdown, m^j, of each bank j is calculated as the average spread over the federal funds rate excluding periods when the federal funds rate is below 1%, as markups/markdowns below are abnormally high/low and bias results. The implied ϵ^j is based on the steady-state relationship between retail and policy rates and calculated as $\epsilon^j = \frac{m^j}{m^j - 1}$. The calibration of three parameters, α^m, $\epsilon^{j,l}$, and $\epsilon^{j,h}$, based on aggregate mean and standard deviation leaves one degree of freedom. I select α^m to target an HHI threshold to minimize distance across moments: unconditional asset-weighted group means and dispersion and distance between model and data group means.

Figure 13 presents comparative statics along extensive, α^m, and intensive margins, ϵ^h for the aggregate deposit and loan rate impact responses to a monetary shock. Panels (a) and (b) show that deposit rate pass-through increases in high-concentration markets' elasticity, $|\epsilon^h|$, and low-concentration markets' share, α^m, and loan rate pass-through decreases in high-concentration markets' elasticity, ϵ^h, and low-concentration markets' share, α^m. Compositional effects play a minor role in low market power environments, i.e., those with high $|\epsilon|$. The U.S. banking sector in 1994 would situate at the back of the loan rate graph and shift to the front over time.

Figure 13: Impact impulse responses to a monetary tightening varying α^m and ϵ^h

(a) Deposit rate

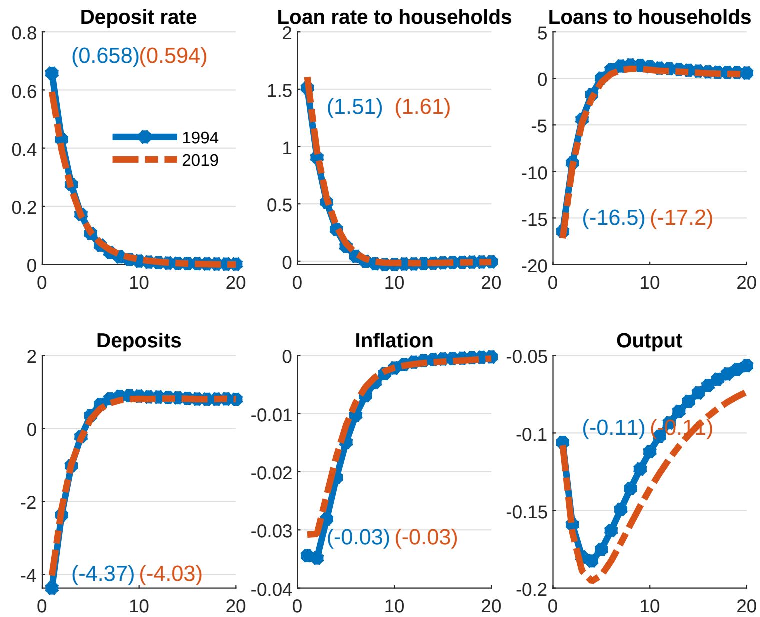
(b) Loan rate



Notes: Aggregate loan and deposit rate impact responses to a monetary shock (z-axis). The y-axis reflects α^m, low-concentration markets' share, the x-axis high-concentration markets' elasticity, ϵ^h, holding ϵ^l constant.

How did U.S. monetary policy transmission change from 1994 to 2019 with more branches located in high-concentration markets, as documented in Figure B.1 in Appendix B.4? Figure 14 contrasts the impulse responses of aggregate deposit and loan rate, deposits, loans, inflation, and output to a monetary shock, with the share of high-concentration markets, $(1 - \alpha^m)$, increasing from 0.3 to 0.6. Focusing first on the loan rate, a comparison of 1994 to 2019 shows that the loan rate was more sensitive to a monetary shock, indicating a greater pass-through in 2019. Loans declined more in response to a policy rate increase, revealing that a larger share of high-concentration markets amplified the credit cycle. In contrast to the loan rate, the deposit rate increased less in 2019, as the banks applied higher markdowns on average. Concentration also affected macroeconomic variables; output contracts slightly more, while inflation decreased less.

Figure 14: Impulse responses to a monetary tightening in 1994 and 2019 varying α^m



Notes: Impulse response functions to a positive monetary shock in 1994 (blue) and 2019 (red). The impact effect is displayed in parentheses.

6.3.2 Heterogeneous Bank Headquarters: Rise in Giant Banks' Share

How do banking sector composition changes affect monetary policy pass-through to deposit and loan rates? The extended model includes two heterogeneous bank types differing in their bank capital ratios, $\nu^{b,j}\ \forall\ j \in \{r, g\}$, labeled regional, r, and giant, g, in line with high and low capital ratios. Appendix B.7 explains the model modifications and includes proofs. Propositions 3 and 4 suggest the aggregate loan rate depends on regional and giant banks' capital ratio, $\nu^{b,r}$ and $\nu^{b,g}$, and regional banks' share, α^b.

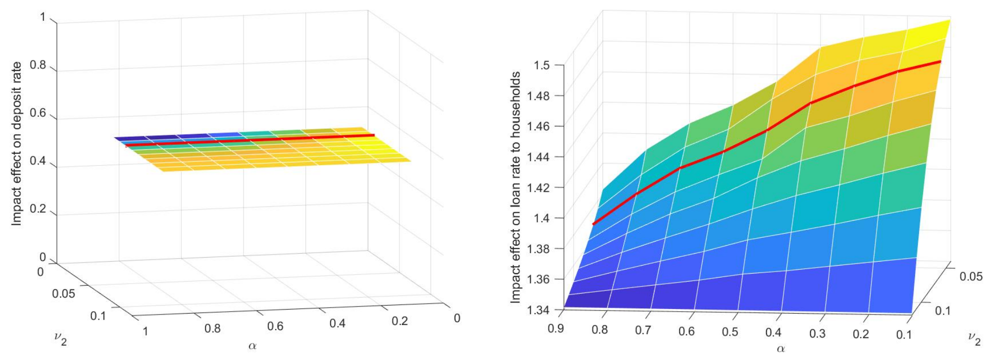
Proposition 3. *Policy rate pass-through to the wholesale funding rates depends inversely on bank capital ratio ν^b; the higher the capital ratio, the less responsive the wholesale funding rate.*

Proposition 4. *Increases in giant banks' market share, $\left(1-\alpha^b\right)$, with a lower bank capital ratio, $\nu^{b,g}$, lead to a higher pass-through, depending on capital ratios' cross-sectional heterogeneity; increases in giant bank's capital ratios, $\nu^{g,r}$, decrease loan rate pass-through.*

Panels (a) and (b) of Figure 15 present aggregate deposit and loan rate impact responses to a monetary shock varying regional banks' share, α^b, and giant banks' capital ratios, $\nu^{b,g}$, holding $\nu^{b,r}$ constant. Loan rate pass-through increases with giant banks' share and decreases in giant banks' capital ratios. Capital ratios do not alter deposit rates, yielding no effect on pass-through.

Figure 15: Impact impulse response to a monetary tightening varying α^b and $\nu^{b,r}$

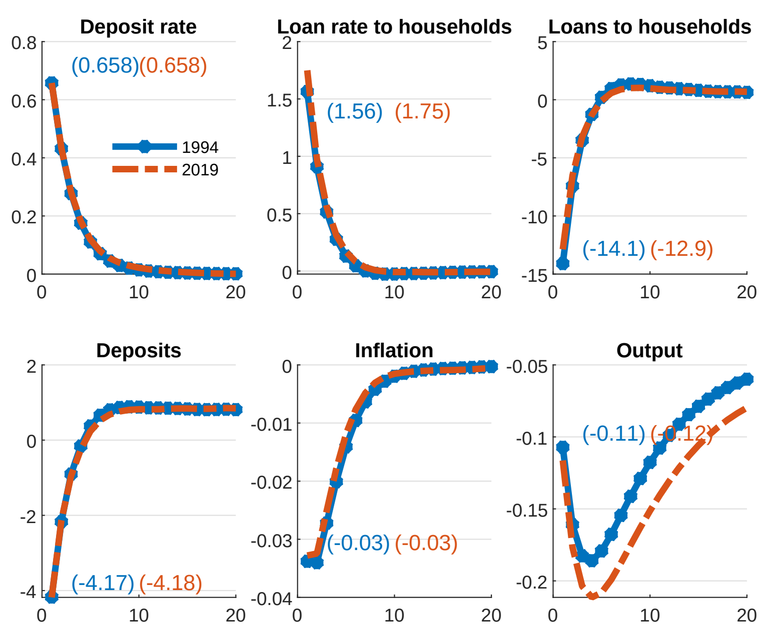
(a) Deposit rate (b) Loan rate



Notes: Aggregate loan and deposit rate impact responses to a positive monetary shock (z-axis). The y-axis corresponds to regional banks' market share, α^b, the x-axis to giant banks' capital ratio, $\nu^{b,g}$, holding the regional banks' capital ratio constant.

How did U.S. monetary policy transmission change from 1994 to 2019 with an increasing share of giant banks? Figure 16 shows the impulse responses to a monetary shock for a giant banks' share, $(1-\alpha^b)$, of 0.1 and 0.6. Comparing 1994 to 2019 reveals that the aggregate loan rate was more responsive to a policy rate increase, while bank size distribution changes again have no impact on the deposit rate. Further, there are effects on the transmission to output and inflation.

Figure 16: Impulse responses to a monetary tightening varying α^b



Notes: Impulse response functions to a positive monetary shock in 1994 (blue) and 2019 (red). The impact effect is displayed in parentheses.

6.3.3 Total Effect of Rise in Bank Concentration

After examining the partial effect of an increasing share of high-concentration markets and giant banks in isolation, I combine both partial effects and consider trends in markups and bank capital ratios over time. Table 6 provides intuition on the expected results:

Table 6: Theoretical predictions on monetary pass-through

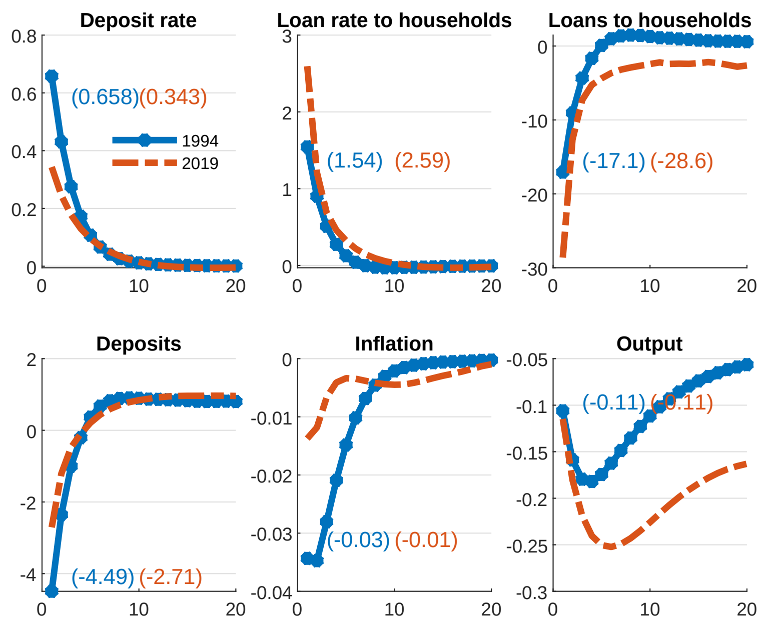
	$\Delta\epsilon^d$	$\Delta\epsilon^l$	$\Delta(1-\alpha^m)$	$\Delta\nu^b$	$\Delta(1-\alpha^b)$
r^l	-	$\uparrow$	$\uparrow$	$\downarrow$	$\uparrow$
r^d	$\downarrow$	-	$\downarrow$	-	-

Notes: Δ stands for change. $\uparrow$ predicts an increase, $\downarrow$ a decrease, and $-$ no change in monetary pass-through.

The comparative statics results suggest rising markups, $\Delta\epsilon^l$, a higher share of high-concentration markets, $\Delta(1-\alpha^m)$, and giant banks, $\Delta(1-\alpha^b)$, increase pass-through to loan rates, with some attenuation from increasing capital ratios, $\Delta\nu^b$. The results also point to a decrease in deposit rate pass-through due to higher markdowns, $\Delta\epsilon^d$, and a higher share of high-concentration markets, $\Delta(1-\alpha^m)$.

Figure 17 shows impulse response functions to a monetary tightening considering changes in α^b, α^m, ϵ, and ν^b, as well as their interaction effects. The results reveal that loan rate pass-through has increased over time, while deposit rate pass-through has declined. Similarly, loans to households declined by more but deposits by less in response to a monetary policy shock. Focusing on macroeconomic variables, transmission to output strengthened, but the effect on inflation dampened. Overall, the differences are more significant than in the partial analysis, suggesting time-varying markups, capital ratios, and interaction effects play a prominent role.

Figure 17: Impulse responses to a monetary tightening varying α^b, α^m, ϵ, and ν^b



Notes: Impulse response functions to a positive monetary shock in 1994 (blue) and 2019 (red). The impact effect is displayed in parentheses.

6.3.4 Rise in Bank Concentration Decomposition

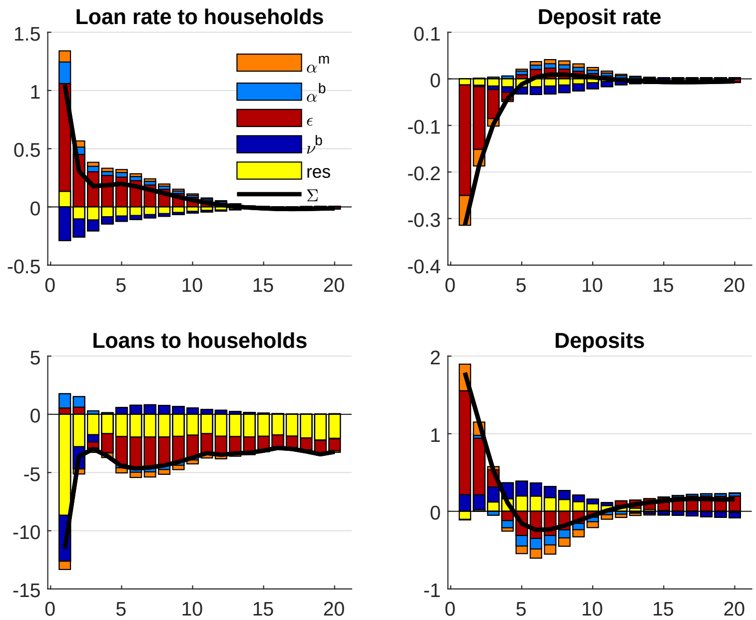
This section decomposes the total effect of rising bank concentration on monetary policy pass-through into five components and compares their relative contribution. In addition to previous literature, the nested model setup also allows assessing interaction effects between the different components, specifically the *market power* and *capital allocation channels*. As summarized in Equation (22), the total effect, Σ, accounts for changes along the extensive and intensive margins. In particular, for changes in: (i) share of low-concentration markets, α^m; (ii) share of regional banks, α^b; (iii) loan demand and deposit supply elasticity, ϵ; (iv) bank capital ratio, ν^b; and (v) an interaction effect, res.

$$\Delta^{\Sigma}_{t+h} = \underbrace{\Delta^{\alpha^m}_{t+h}}_{\substack{\text{\% high-}\\ \text{concentration}\\ \text{markets}}} + \underbrace{\Delta^{\alpha^b}_{t+h}}_{\text{\% small banks}} + \underbrace{\Delta^{\epsilon}_{t+h}}_{\text{markup}} + \underbrace{\Delta^{\nu^b}_{t+h}}_{\text{bank capital ratio}} + \underbrace{res_{t+h}}_{\text{interaction}}, \tag{22}$$

where $\Delta^j_{t+h}\ \forall\ j \in \{\Sigma, \alpha^m, \alpha^b, \epsilon, \nu^b, res\}$ reflects the difference between the impulse response functions of each variable from 2019 and 1994 under calibration j, calculated as $\Delta^j_{t+h} = IRF^{j,2019}_{t+h} - IRF^{j,1994}_{t+h}$ for each horizon.[38]

[38] The equation omits superscripts h for readability. The difference in impulse response functions is expressed in levels for interest rates and in percentage point deviations for all other variables. The interaction

Figure 18: Decomposing the change in monetary pass-through to rates and volumes

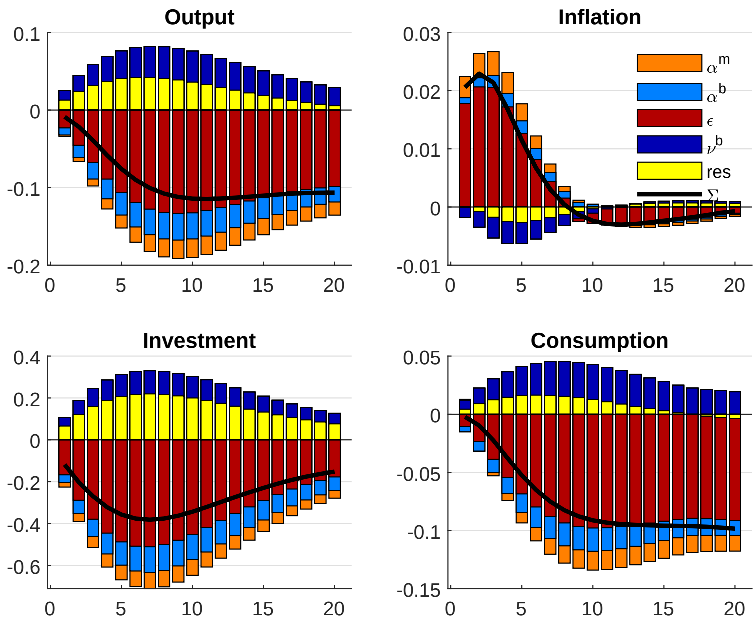


Notes: Decomposition of total effect, Σ, into five components: changes in share of low-concentration markets, α^m, share of regional banks, α^b, elasticity of loan demand and deposit supply, ϵ, bank capital ratio, ν^b, and an interaction effect, res. The x-axis represents the horizon.

Figure 18 decomposes the total change in monetary policy pass-through for the aggregate deposit rate, loan rate, household loans, and deposits. Increasing markups, ϵ primarily drive the total increase in loan rate pass-through, and to some degree, compositional shifts in α^m and α^b, and interaction effects, res. Bank capital ratios, ν^b, have a negative impact. Pass-through to the deposit rate declined due to increasing markdowns from shifts along the intensive and extensive margins (i.e., increases in α^m and decreases in $|\epsilon|$). Aggregate loans and deposits present a near mirror image of the aggregate loan and deposit rate, with the decrease compromised predominantly of ϵ, but interaction effects play a more significant role. Rising markups interact with financial frictions and lead to a more substantial decline in lending; a partial analysis would fail to capture that. Overall, the importance of markup shifts indicates that secular trends outweigh composition effects.

effect equals the residual.

Figure 19: Decomposing the change in monetary policy transmission to the macroeconomy



Notes: Total effect, Σ, is decomposed into five components: changes in share of low-concentration markets, α^m, and regional banks, α^b, loan demand and deposit supply elasticity, ϵ, bank capital ratio, ν^b, and an interaction effect, res. The x-axis represents the horizon.

Figure 19 presents the decomposition for macroeconomic variables. Recall from Figure 17 that total monetary policy transmission to output, investment, and consumption strengthened in 2019; that is, those variables declined more in response to a positive shock, also reflected by the negative difference. Figure 19 reveals that the amplification results mostly from rising markups, ϵ. The rise in bank capital ratios, ν^b, and interaction effects, res, counteracted the amplification. Hence, a partial analysis would overstate the total effects. In contrast to output, monetary policy transmission to inflation is more muted, indicating that rising bank concentration has opposite implications for the transmission to prices and output.

6.3.5 Implications on the Phillips Curve

To examine the impact on the slope of the Phillips curve, I derive the model's log-linearized Phillips curve, expressing changes in current inflation, $\widetilde{\pi}_t$, in terms of changes in output, $\widetilde{y}_t$, and expected future inflation, $\mathbb{E}_t\widetilde{\pi}_{t+1}$ (starting from equation (50) in Appendix B.2):[39]

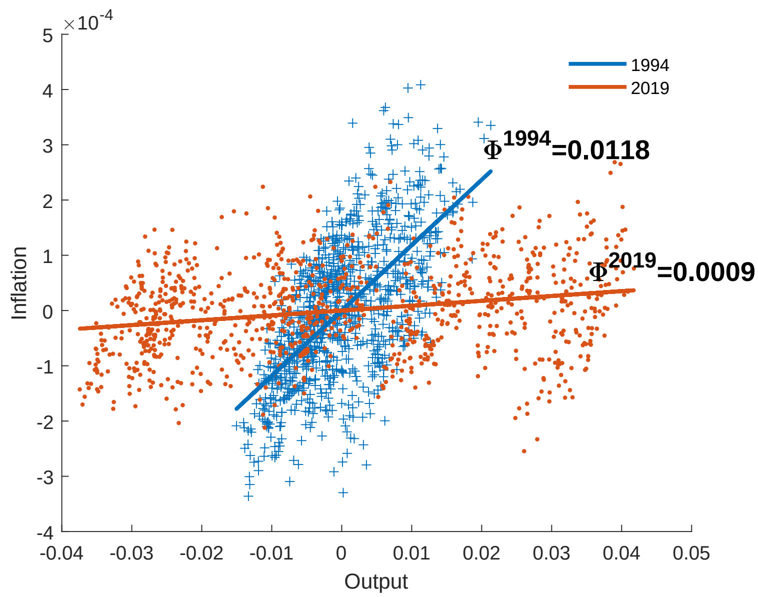
$$\widetilde{\pi}_t = \Phi\widetilde{y}_t + \beta^P\mathbb{E}_t\widetilde{\pi}_{t+1}, \tag{23}$$

[39]Equation (23) abstracts from indexation, $\iota_p = 0$; not simulation.

where Φ summarizes the coefficients in front of output, such as Rotemberg price adjustment, κ_p, and elasticity of substitution across goods, ϵ^y.

Figure 20 shows the inflation-output relationship based on simulated data for the 1994 and 2019 model calibrations.[40] Table 5 presents the calibration details. The simulation is based on 1,000 periods and includes 10,000 initial burn-in periods. The monetary shock is the only source of stochastic uncertainty.[41] A comparison of the two calibrations' estimated slope indicates the Phillips curve flattens over time, consistent with recent empirical evidence. For example, Hazell et al. (2020) study the relationship between inflation and unemployment across U.S. states for the periods 1978-1990 and 1991-2018 and find that the Phillips curve flattens by a factor of 2 to 100, depending on model specification. My calibration reveals a decline by a factor of 13.

Figure 20: Phillips curves: relation between inflation and output



Notes: Simulated data for output and inflation based on banking sector calibration to 1994 and 2019. Data expressed in terms of deviations from the steady-state level (unconditional mean).

What is the mechanism behind the flattening of the Phillips curve? The result relies upon two sets of factors. First, output and the slope of the Phillips curve depends on the level of resource costs responsible for a wealth effect. In simple words, the financial sector absorbs part of the resources in the economy, such as operating costs bank management

[40]To control for changes in inflation expectations, the y-axis shows: $\pi_t - \beta\mathbb{E}_t\pi_{t+1}$. Alternatively, controlling for inflation expectations in the regression yields similar results.

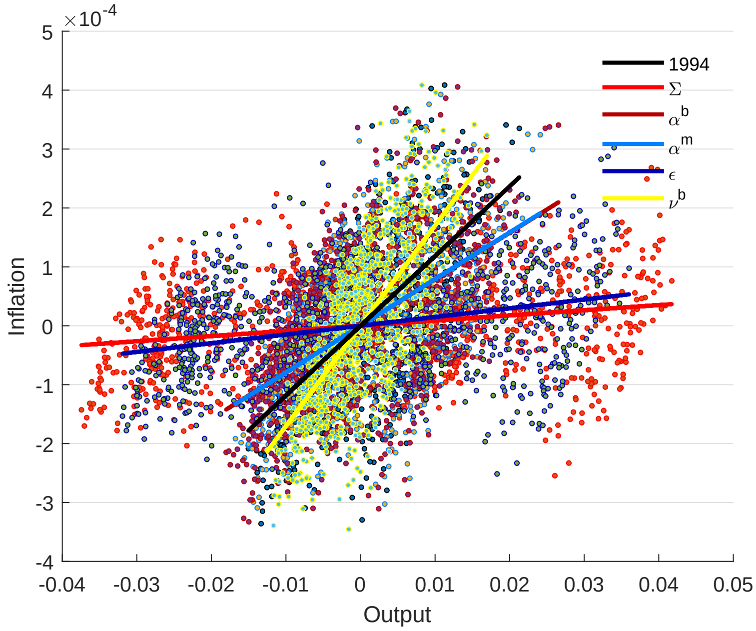
[41]The results hold for any demand shock (e.g., monetary shock, preference shock, housing preference shock). Supply shocks cause the Phillips curve to shift and are not suitable for estimating the Phillips curve's slope.

costs (reflected by δ^b in Equation (8)), and creates a dead-weight loss.[42] With rising bank concentration and higher bank management costs, "effective" output (i.e., output net off adjustment and management costs), becomes more volatile and disentangles from production. Second, the slope of the Phillips curve depends on the level of frictions affecting labor supply. Wage rigidities and habit formation interact with the wealth channel, further breaking the link between output, marginal costs, and inflation, see also Appendix B.8 for more details. Wider wedges and markups affect the stochastic discount factor of savers and borrowers and the inter-temporal consumption and labor supply decision. While a higher loan rate and stochastic discount factor of borrowers lead to a stronger response today, lower saving rates and a higher stochastic discount factor of savers make future consumption more valuable and lead to a lower response today. Overall the first effect outweighs and net consumption contracts by more. Note that the result is robust to redefining output net of resource costs, but the flattening decreases. Similarly, the result does not go away by eliminating labor market frictions but reduces the magnitude of the effect. However, the effect disappears by redefining output and eliminating all labor supply frictions. I regard my results on the flattening of the Phillips Curve as complementary to existing explanations that point to changes in the conduct of monetary policy and inflation expectations (e.g., Carlstrom et al., 2009), less frequent price adjustments (e.g., Kuttner and Robinson, 2010), and higher worker bargaining power (e.g., Ng et al., 2018).

What is the relative importance of the market power and capital allocation channels for the Phillips curve flattening? I analyze the marginal impact of structural changes in (i) low-concentration markets' share, α^m, (ii) regional banks' share, α^b, (iii) loan demand and deposit supply elasticity, ϵ, and (iv) bank capital ratio, ν^b. Figure 21 contrasts the estimated Phillips curves for each specification with the 1994 baseline. I find that rising markups, ϵ, are the main driver. Although changes along the extensive margin, market shares of regional banks, α^b, and low-concentration markets, α^m, shift the Phillips curve in the same direction, their effects are relatively small. An increase in bank capital ratios, ν^b, leads to a steeper curve, slightly counteracting the other forces. The findings are consistent with the decomposition in Section 6.3.4, and confirm the relevance of the *market power channel*.

[42] Redistributing the management costs to the patient household as transfer still yields qualitatively similar results and a flattening of the Phillips curve over time.

Figure 21: Phillips curves based on different calibrations



Notes: Simulated data based on different banking sector calibrations. 1994 reflects the baseline calibration. Σ considers all structural changes, including changes in regional banks' share, α^b, low-concentration markets' share, α^m, demand elasticity, ϵ, and bank capital ratio, ν^b. Data is expressed in terms of deviations from the steady-state level (i.e., unconditional mean).

7 Conclusion

This paper examines how the banking sector's structure affects monetary policy pass-through at a disaggregated level. I suggest that it is essential to look at observed differences in retail deposit and loan rates within a given bank across regions and bank institutions within a region. The variation in retail rates sheds light on how the composition of local markets and the size distribution of banks affect the aggregate transmission of monetary policy via two channels: a *market power channel*, with higher concentration in local banking markets increasing spreads and markups, and a *capital allocation channel*, with higher banking concentration implying a lower aggregate banking sector capitalization and amplifying financial frictions due to regulation.

I deliver theoretical and empirical evidence for the heterogeneous monetary policy pass-through to loan and deposit rates in the cross-section and over time and incorporate that into a quantitative model. I explain the cross-sectional heterogeneity via differences in market power across locations and marginal costs across banks stemming from bank capital ratios and time-series variation with asymmetric adjustment costs for expanding the lending volume. Counterfactual analyses in a New Keynesian model with heterogeneous bank branches and banks calibrated to the 1994 and 2019 reveal that the rise in bank concentration strengthened monetary policy pass-through to loan rates and amplified the credit cycle. I decompose the effect on pass-through and find that both increased market power and banks' size distribution changes, as well as their interactions, amplified monetary policy pass-through. The rise in bank concentration amplifies monetary policy transmission to output and investment but dampens its impact on inflation. The opposing effects lead to a flattening of the Phillips curve over time.

This paper suggests that rising bank concentration has important implications for monetary policy transmission and effectiveness. For the conduct of optimal monetary policy, both market power and capitalization of banks should be taken into account and considered jointly. The results indicate that monetary policy became more potent over time. In other words, nowadays, the central bank needs to adjust the policy rate by less to achieve a similar effect on output, though there are limited effects on inflation. Further, the results inform about heterogeneity at a disaggregated level for policy design. Future work could expand the model to heterogeneous banks of more than two types and locations and closely study distributional effects across U.S. counties and disparities.

References

Abbritti, M. and Fahr, S. (2013). Downward wage rigidity and business cycle asymmetries. *Journal of Monetary Economics*, 60(7):871–886.

Allen, J., Clark, R., and Houde, J.-F. (2019). Search frictions and market power in negotiated-price markets. *Journal of Political Economy*, 127(4):1550–1598.

Altavilla, C., Canova, F., and Ciccarelli, M. (2019). Mending the broken link: Heterogeneous bank lending rates and monetary policy pass-through. *Journal of Monetary Economics*.

Andres, J. and Arce, O. (2012). Banking competition, housing prices and macroeconomic stability. *The Economic Journal*, 122(565):1346–1372.

Ball, L. M. and Mazumder, S. (2011). Inflation dynamics and the great recession. Technical report, National Bureau of Economic Research.

Berger, A. N. and Hannan, T. H. (1989). The price-concentration relationship in banking. *The Review of Economics and Statistics*, 71(2):291–299.

Bluedorn, J. C., Bowdler, C., and Koch, C. (2017). Heterogeneous bank lending responses to monetary policy: New evidence from a real-time identification. *International Journal of Central Banking*, 47:95–149.

Borenstein, S., Cameron, A. C., and Gilbert, R. (1997). Do gasoline prices respond asymmetrically to crude oil price changes? *Quarterly Journal of Economics*, 112(1):305–339.

Brennecke, C., Jacewitz, S., and Pogach, J. (2021). Shared destinies? Small banks and small business consolidation. *Small Banks and Small Business Consolidation. Federal Reserve Bank of Kansas City Working Paper*, (21-19).

Brunnermeier, M. K. and Koby, Y. (2018). The reversal interest rate. Technical report, National Bureau of Economic Research.

Carlstrom, C. T., Fuerst, T. S., and Paustian, M. (2009). Monetary policy shocks, Choleski identification, and DNK models. *Journal of Monetary Economics*, 56(7):1014–1021.

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

De Loecker, J., Eeckhout, J., and Unger, G. (2020). The rise of market power and the macroeconomic implications. *The Quarterly Journal of Economics*, 135(2):561–644.

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

Drechsler, I., Savov, A., and Schnabl, P. (2018). Banking on deposits: Maturity transformation without interest rate risk. Technical report, National Bureau of Economic Research.

Driscoll, J. C. and Judson, R. (2013). Sticky deposit rates. *Available at SSRN 2241531*.

Fahr, S. and Smets, F. (2010). Downward wage rigidities and optimal monetary policy in a monetary union. *Scandinavian Journal of Economics*, 112(4):812–840.

Flannery, M. J. (1982). Retail bank deposits as quasi-fixed factors of production. *The American Economic Review*, 72(3):527–536.

Freixas, X. and Rochet, J.-C. (2008). *Microeconomics of banking*. MIT press.

Gerali, A., Neri, S., Sessa, L., and Signoretti, F. M. (2010). Credit and banking in a DSGE Model of the euro area. *Journal of Money, Credit and Banking*, 42:107–141.

Gertler, M. and Karadi, P. (2015). Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics*, 7(1):44–76.

Grant, C. (2007). Estimating credit constraints among US households. *Oxford Economic Papers*, 59(4):583–605.

Hazell, J., Herreño, J., Nakamura, E., and Steinsson, J. (2020). The slope of the Phillips Curve: evidence from US states. Technical report, National Bureau of Economic Research.

Iacoviello, M. (2005). House prices, borrowing constraints, and monetary policy in the business cycle. *American Economic Review*, 95(3):739–764.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, 95(1):161–182.

Kashyap, A. K. and Stein, J. C. (2000). What do a million observations on banks say about the transmission of monetary policy? *American Economic Review*, 90(3):407–428.

Kishan, R. P. and Opiela, T. P. (2000). Bank size, bank capital, and the bank lending channel. *Journal of Money, Credit, and Banking*, 32(1):121.

Klein, M. A. (1971). A theory of the banking firm. *Journal of Money, Credit and Banking*, 3(2):205–218.

Kuttner, K. and Robinson, T. (2010). Understanding the flattening Phillips curve. *The North American Journal of Economics and Finance*, 21(2):110–125.

Levieuge, G. and Sahuc, J.-G. (2021). Downward interest rate rigidity. *European Economic Review*, 137:103787.

Matheson, T. and Stavrev, E. (2013). The great recession and the inflation puzzle. *Economics Letters*, 120(3):468–472.

Meyer, A. P. (2018). Market concentration and its impact on community banks. *St. Louis Fed or Federal Reserve System*.

Monti, M. (1972). *Deposit, credit and interest rate determination under alternative bank objective function*. North-Holland/American Elsevier.

Nakamura, E. and Steinsson, J. (2018). High-frequency identification of monetary non-neutrality: the information effect. *Quarterly Journal of Economics*, 133(3):1283–1330.

Neumark, D. and Sharpe, S. A. (1992). Market structure and the nature of price rigidity: evidence from the market for consumer deposits. *The Quarterly Journal of Economics*, 107(2):657–680.

Ng, M., Wessel, D., and Sheiner, L. (2018). The Hutchins Center explains: The Phillips curve. *Brookings Up Front*.

Peltzman, S. (2000). Prices rise faster than they fall. *Journal of Political Economy*, 108(3):466–502.

Ramey, V. A. and Zubairy, S. (2018). Government spending multipliers in good times and in bad: evidence from us historical data. *Journal of Political Economy*, 126(2):850–901.

Romer, C. D. and Romer, D. H. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, 94(4):1055–1084.

Scharfstein, D. and Sunderam, A. (2016). Market power in mortgage lending and the transmission of monetary policy. *Working Paper*.

Ulate, M. (2021). Going negative at the zero lower bound: The effects of negative nominal interest rates. *American Economic Review*, 111(1):1–40.

Van den Heuvel, S. J. (2002). The bank capital channel of monetary policy. *The Wharton School, University of Pennsylvania, mimeo*, pages 2013–14.

Wang, O. (2019). Banks, low interest rates, and monetary policy transmission. Working paper, MIT.

Wang, Y., Whited, T. M., Wu, Y., and Xiao, K. (2018). Bank market power and monetary policy transmission: Evidence from a structural estimation. *Available at SSRN 3049665*.

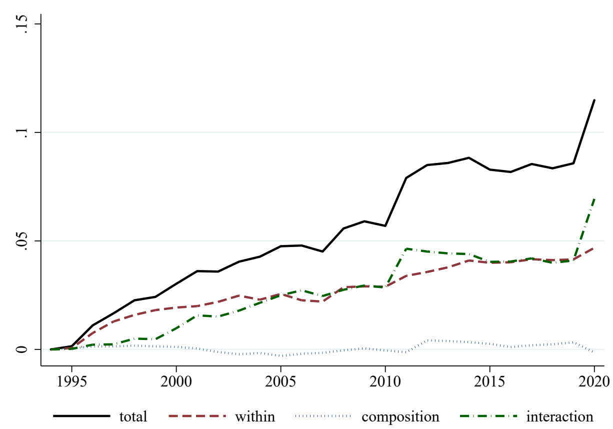
Yankov, V. (2014). In search of a risk-free asset. Technical report, Board of Governors of the Federal Reserve System (US).

A Empirical Appendix

A.1 Decomposition of Rise in U.S. Bank Concentration

To what extent is the rise in bank concentration a general trend seen in all U.S. counties or driven by compositional effects? The increase in aggregate national bank concentration consists of three parts: (i) changes in concentrated counties' relative market size, (ii) changes in within-county bank concentration, and (iii) interaction effects.[43] Figure A.1 shows that the main drivers are increases within-county and the interaction effect, contributing 0.05 and 0.07, respectively, to the total increase of 0.11 from 1994 to 2020.

Figure A.1: Decomposition of rise in U.S. HHI



Notes: Decomposition of national HHI growth from Figure 1(a) in: (i) changes in share of high-concentration counties, (ii) changes in concentration within-county, and (iii) interaction effects.

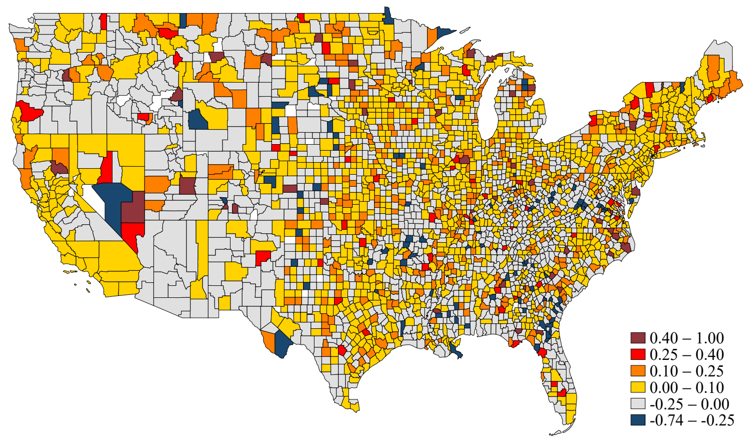
Figure A.2 takes a close look at the changes in local bank concentration between 1994 and 2019. While many rural counties observed a decrease in concentration, urban counties with a relatively larger banking sector in terms of total deposits observed an increase in local bank concentration. The increase in concentration in large counties drives the national pattern in Figure 1(a). Shifts across time are also driven by allocating assets to headquarters in NYC and Salt Lake City. With this in mind, it is easy to reconcile that an analysis of the simple average of local HHIs or the number of banks operating in a county shows a slightly different pattern but likewise a renewed increase in concentration (Brennecke et al., 2021; Meyer, 2018).

[43] Decomposition of the cumulative growth in national HHI relative to 1994:

$$HHI_t - HHI_{1994} = \sum_c \left\{ \underbrace{d^c_{94}\left(HHI^c_t - HHI^c_{1994}\right)}_{\text{within}} + \underbrace{HHI_{1994}\left(d^c_t - d^c_{1994}\right)}_{\text{composition}} + \underbrace{\left(d^c_t - d^c_{1994}\right)\left(HHI^c_t - HHI^c_{1994}\right)}_{\text{interaction}} \right\}$$

where HHI^c_t and d^c_t are the HHI and deposit market share of county c. The first term on the right-hand side reflects shifts within-county, the second term the share shift, and the last term interaction effects.

Figure A.2: Change in bank concentration between 1994 and 2019 by county

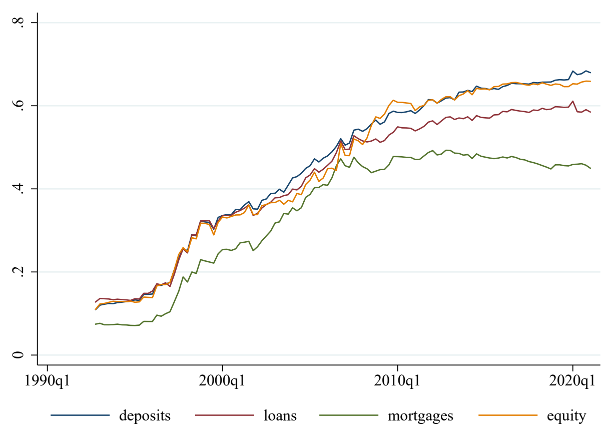


Notes: Changes in HHIs between 1994 and 2019: $HHI_c^{2019} - HHI_c^{1994}$. Source: FDIC Summary of Deposits.

A.2 Alternative Concentration Measures

The benchmark concentration measurement of this paper is based on deposit market shares as there is no information on other balance sheet items at the *branch* level. This is arguably not a major caveat, since concentration measures for deposit, credit, and mortgage markets are highly correlated at a national level, and all have increased over the past two decades (shown in Figure A.3).

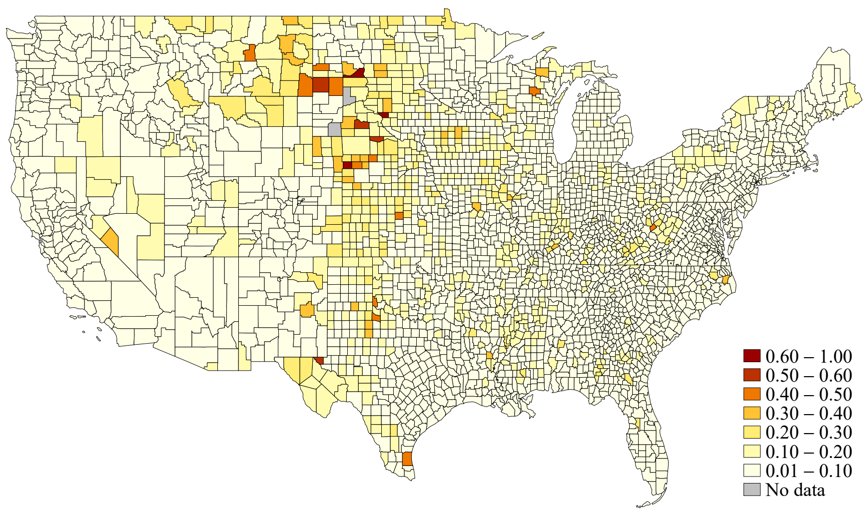
Figure A.3: Percent of different assets held by giant banks



Notes: Market share of giant banks based on deposits, loans, real estate loans and equity. Source: Federal Deposit Insurance Corporation.

In addition, the Home Mortgage Disclosure Act (HMDA) provides data on mortgage originations at a county level, a close substitute relying on flows instead of stocks. Following Scharfstein and Sunderam (2016), mortgage market concentration is based on each institution's share of mortgage originations per county/year. In contrast to the FDIC data, HMDA includes non-bank lenders and credit unions. Figure A.4 shows a map of the mortgage market concentration across counties in 2019. Overall, the mortgage market concentration is much lower than the deposit market concentration reflected by lighter coloring. Similar to deposit market concentration shown in Figure 2, high mortgage market concentration is predominantly in the Midwest and center of the United States. Table A.1 quantifies the relationship between the concentration measures for different periods. The correlation is relatively strong, consistently about .42 across time. Appendix A.6 presents the pass-through results conditional on mortgage market concentration.

Figure A.4: Mortgage origination concentration by county based on HMDA data



Notes: County-level HHIs based on mortgage originations in 2019. Source: Home Mortgage Disclosure Act.

Table A.1: Correlation of county-level deposit and mortgage market concentration

	2000-2019	2000-2008	2009-2019
$\rho(HHI^{dep}, HHI^{mortg})$	0.42	0.42	0.43

Notes: ρ reflects the correlation coefficient of the county-level HHIs based on deposits and mortgage originations. Source: FDIC Summary of Deposits, Home Mortgage Disclosure Act.

A.3 Survey Instrument

Figure A.5 shows a template of the *RateWatch* survey instrument. This survey is sent out to branch loan officers on a monthly basis to collect information on prices for financial advisors and conduct competitor analyses for individual clients. *RateWatch* collects offered loan rate quotes to the "best" customer, i.e. clients with the excellent credit scores. To obtain standardized loan rates across branches and time, *RateWatch* asks for offered rates with close to zero fees and points, and a constant loan amount, e.g. a 30-year mortgage rates with a loan amount of $175,000.

Figure A.5: Survey instrument

Institution Name:
Account Number:
Contact:
Today's Date:

Current Prime Rate:

Send to: submitrates@rate-watch.com

RATEWATCH
Accurate *Financial Data* Since 1989

RATEWATCH PHONE 800.348.1831

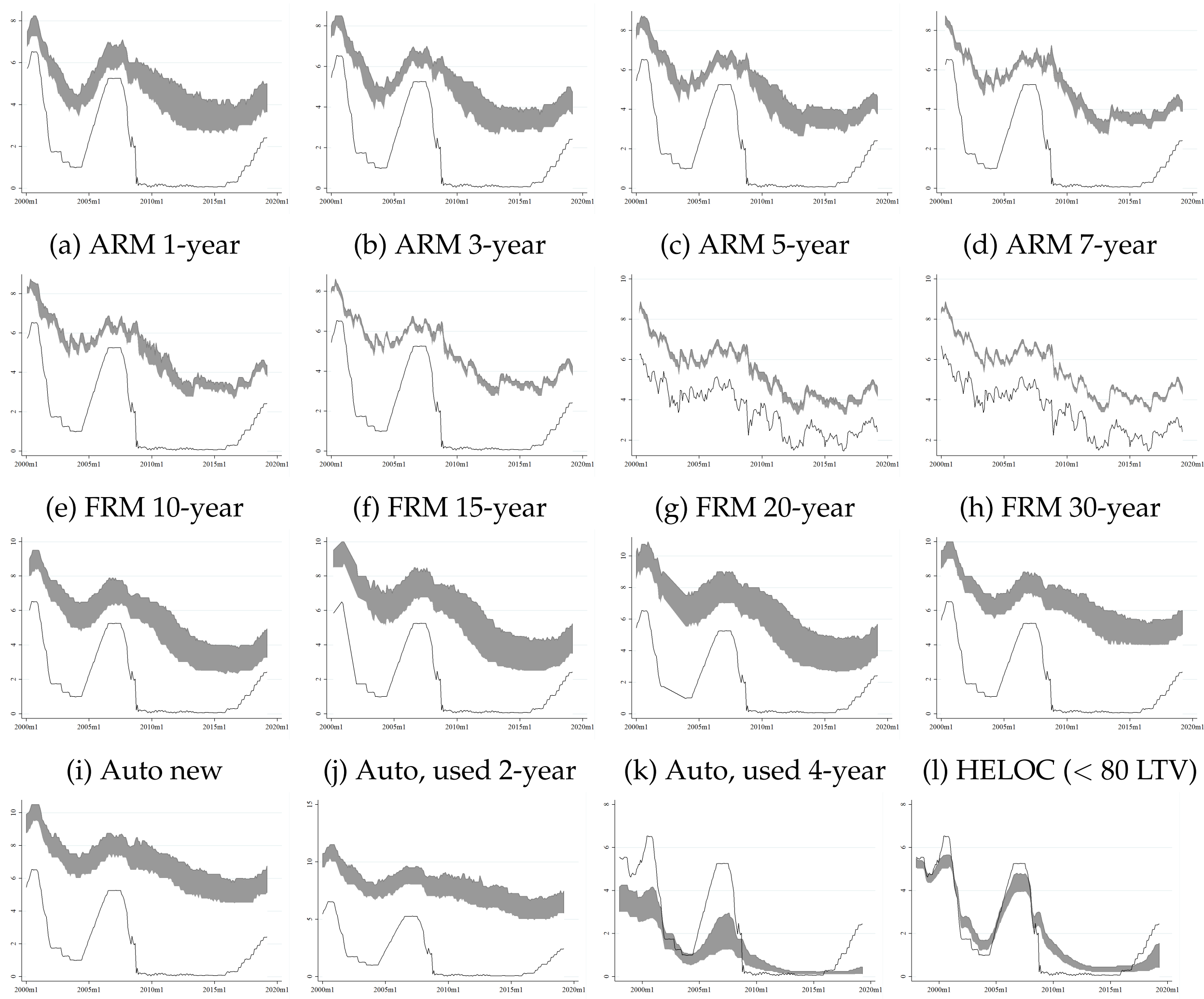
Mortgages: Please list in-house rates first. If N/A then 2nd market rates. If not offered then N/A the category. Need as close to zero point/fees @ 60 day lock period. Purchase, single family owner occupied.

1 YEAR ARM @ 175K LOAN AMOUNT	FIXED RATE	COMMENTS
RATE		
APR		
DISCOUNT POINTS		
DOWN PAYMENT TO AVOID PMI		
CAPS		
MAX AMORTIZATION TERM		
ORIGINATION FEES		

3 YEAR ARM @ 175K LOAN AMOUNT	FIXED RATE	COMMENTS
RATE		
APR		
DISCOUNT POINTS		
DOWN PAYMENT TO AVOID PMI		
CAPS		
MAX AMORTIZATION TERM		
ORIGINATION FEES		

Notes: A template of the survey instrument *RateWatch* sends out to bank branches. Source: RateWatch.

A.4 Dispersion and Spread Over Time



(a) ARM 1-year (b) ARM 3-year (c) ARM 5-year (d) ARM 7-year

(e) FRM 10-year (f) FRM 15-year (g) FRM 20-year (h) FRM 30-year

(i) Auto new (j) Auto, used 2-year (k) Auto, used 4-year (l) HELOC (< 80 LTV)

(m) HELOC (80-90) (n) HELOC (> 90 LTV) (o) Saving rates (p) Time deposits

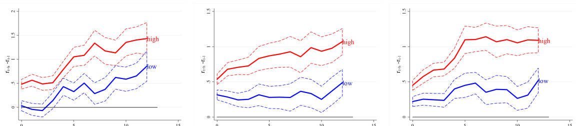
Notes: IQR of branch-level deposit and loan rates. ARM denotes adjustable rate mortgage, FRM, fixed rate mortgage, with a loan amount of $ 175,000 and maturity of 30 years. HELOC stands for home equity line of credit rates with varying loan-to-value (LTV) ratios. Auto loan rates vary by car age (36 months contracts).

A.5 Alternative Monetary Shocks

This robustness check contrasts the results using Nakamura and Steinsson (2018) monetary shocks (Baseline), based on the first principal component of surprise movements in five futures, to surprises in the current month's future rate ($MP1$), three-month ahead future rate ($FF4$),[44] Romer and Romer (2004) narrative monetary shocks ($R\&R$), [45] and raw changes in the federal funds rate (dFF_t).

Figure A.7: Bank concentration

(a) Baseline (b) $MP1$ (c) $FF4$



(d) $R\&R$ (e) dFF_t

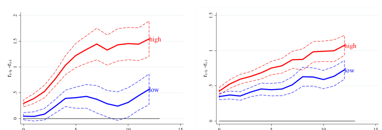
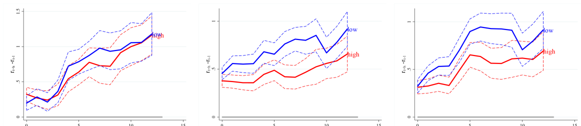


Figure A.8: Bank capitalization

(a) Baseline (b) $MP1$ (c) $FF4$



(d) $R\&R$ (e) dFF_t

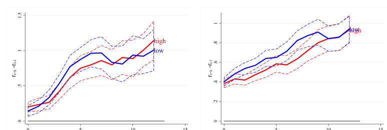
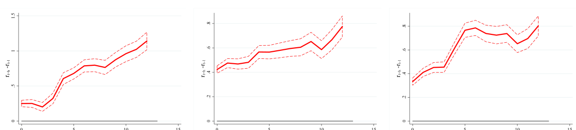


Figure A.9: Linear

(a) Baseline (b) $MP1$ (c) $FF4$



(d) $R\&R$ (e) dFF_t

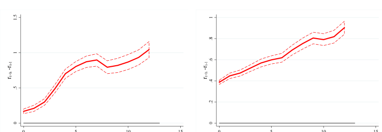
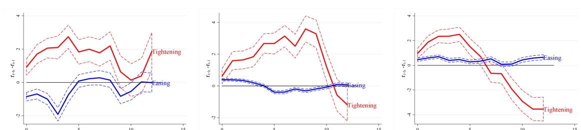
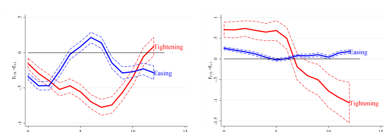


Figure A.10: Asymmetries

(a) Baseline (b) $MP1$ (c) $FF4$



(d) $R\&R$ (e) dFF_t



Notes: Impulse response functions of 1-year hybrid ARM rates to a monetary policy shock at both high and low local bank concentrations in Figure A.7, at both high and low local bank capitalization in Figure A.8, unconditionally in Figure A.9, and at monetary tightening and easing periods in Figure A.10. Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

Figures A.7 to A.8 reveal that the general pattern holds across monetary policy shocks and using changes in the federal funds rate. In addition, Figure A.10 shows asymmetries between monetary easing and tightening across monetary policy shocks.

[44]Gertler and Karadi (2015) use the three-month ahead future surprise to identify monetary shocks.

[45]Romer and Romer (2004)'s shock series ends in 1996. I use an updated shock series provided by Johannes Wieland for the period 2000-2007.

A.6 HMDA Mortgage Market Concentration

Appendix A.6 compares the importance of deposit and mortgage market concentration for monetary policy pass-through to loan rates. Panels (a) to (e) in Figures A.11 and A.12 show the conditional loan rate responses to different monetary shocks: (i) Nakamura and Steinsson (2018) surprises, (ii) current month's future rate ($MP1$) surprises, (iii) three-month ahead future rate ($FF4$) surprises, (iv) Romer and Romer (2004) narrative monetary shocks ($R\&R$), and (v) raw changes in the federal funds rate (dFF_t).

Figure A.11: Deposit market concentration

(a) Baseline (b) $MP1$ (c) $FF4$ (d) $R\&R$ (e) dFF_t

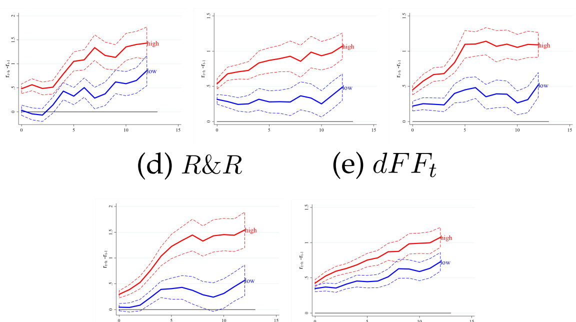
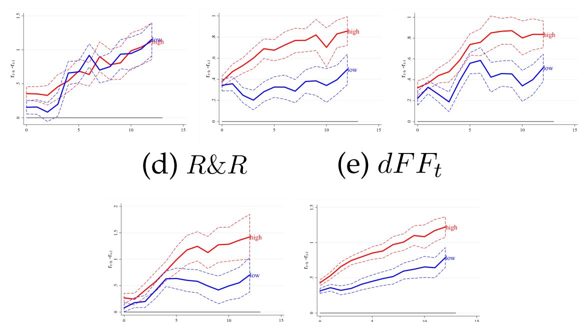


Figure A.12: Mortgage market concentration

(a) Baseline (b) $MP1$ (c) $FF4$ (d) $R\&R$ (e) dFF_t

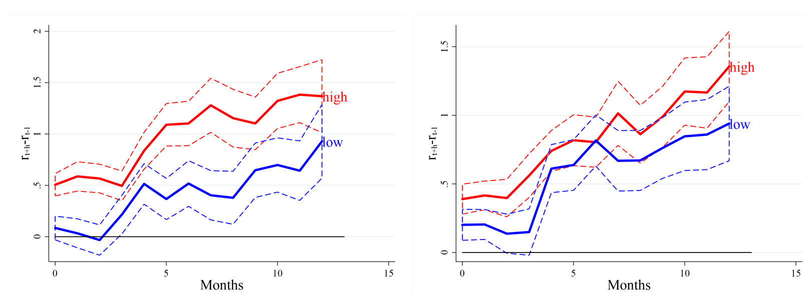


Notes: Impulse responses of 1-year hybrid ARM rates to a monetary shock at both high and low concentration levels. Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

Figures A.11 and A.12 reveal that both concentration measures correlate positively with pass-through. Bank branches operating in markets with high deposit and mortgage concentration adjust loan rates by more and faster to a monetary shock. Quantitatively, deposit market power plays a larger role in loan rate pass-through, reflected by a larger difference. A horse race between the two, estimating a regression model with both concentration measures jointly, indicates that both are important independently (Figure A.13).

Figure A.13: Including both concentration measures simultaneously

(a) Deposit market (b) Mortgage market



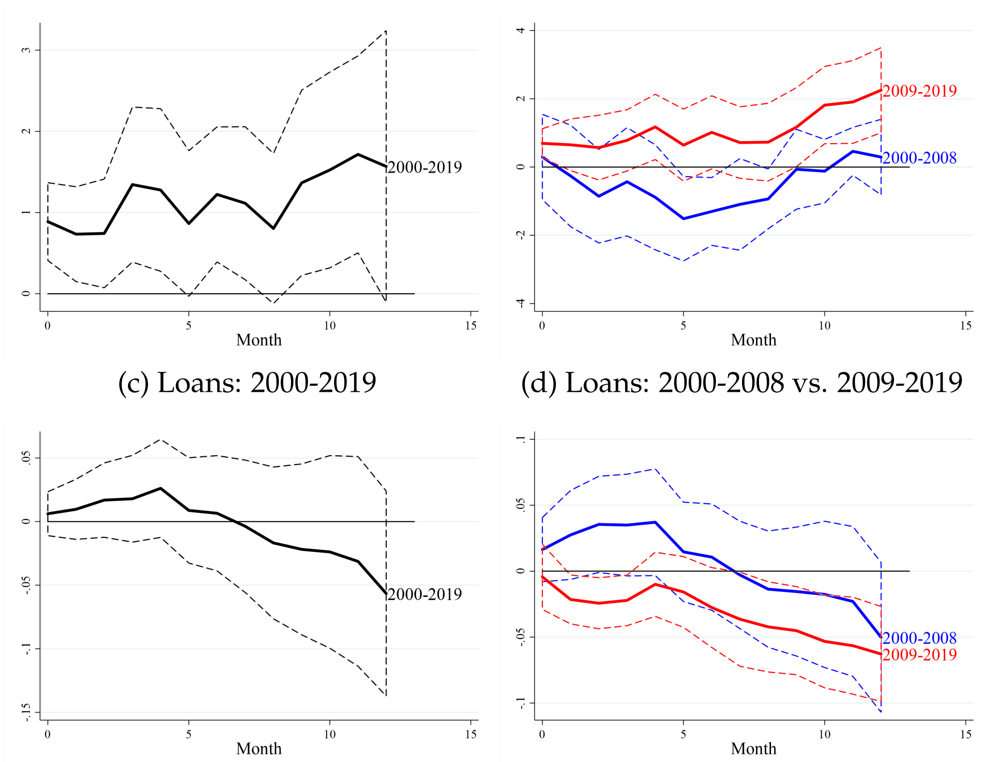
Notes: Impulse responses of 1-year hybrid ARM rates to a monetary shock at both high and low concentration levels. Horizon is in months, and standard errors are clustered at the county level (90% confidence intervals).

A.7 Increasing Aggregate Pass-Through Over Time

Figure A.14 presents responses of the aggregate mortgage rate and real estate loans to Nakamura and Steinsson (2018) monetary policy surprises for the entire and a center-split sample, controlling for aggregate national economic and financial conditions by interacting these with the monetary policy surprises. The aggregate pass-through to mortgage rates is higher in the second period (red) than in the first period (blue). Similarly, real estate lending declined more sharply in the second period. The increasing aggregate pass-through between 2000-2008 vs. 2009-2019 is in line with the expected increase due to compositional shifts in the banking sector and bank concentration over the same time.

Figure A.14: Impulse responses of aggregate mortgage rates and real estate lending by period

(a) Rate: 2000-2019 (b) Rate: 2000-2008 vs. 2009-2019



Notes: Impulse response functions of the aggregate mortgage rate and log real estate loans from commercial banks to a monetary policy shock.

A.8 State-Dependent Monetary Policy Pass-Through

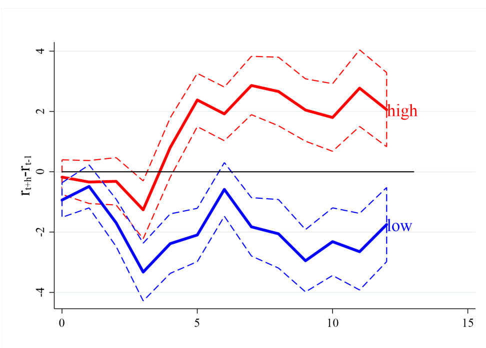
This section documents the results for double interaction terms confirming that the relation between pass-through and bank concentration and capitalization holds across states. I regress branch i's rate adjustment, $r^l_{t+h,i,c} - r^l_{t-1,i,c}$, on monetary shock, s_t, interacted with bank concentration or capitalization, and an indicator for expected monetary tightening, $\mathbb{I}(\mathbb{E}_{t-1}\Delta r^f_t > 0)$, and easing, $\mathbb{I}(\mathbb{E}_{t-1}\Delta r^f_t < 0)$:

$$r^l_{t+h,i,c} - r^l_{t-1,i,c} = \alpha^h_i + \beta^h s_t + \mathbb{I}\left(\mathbb{E}_{t-1}\Delta r^f_t > 0\right)\left(\alpha^{h,+}_i + \beta^{h,+} s_t + \gamma^{h,+} s_t \times X_{t-1,i,c}\right) + \mathbb{I}\left(\mathbb{E}_{t-1}\Delta r^f_t < 0\right)\left(\alpha^{h,-} + \beta^{h,-} s_t + \gamma^{h,-} s_t \times X_{t-1,i,c}\right) + \theta^h X_{t-1,i,c} + \eta^h Z_{t,c} + \epsilon_{t+h,i,c} \quad (24)$$

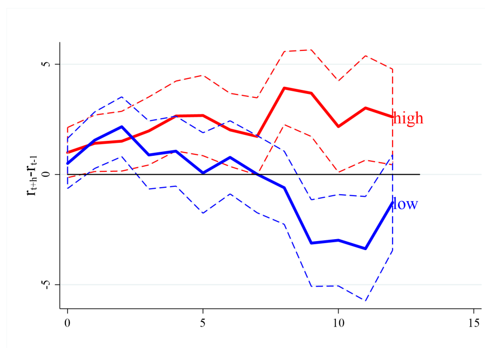
Figures A.15 shows loan rate impulse responses separately for monetary tightening and easing, differing by the level of bank concentration or capitalization.

Figure A.15: Impulse responses of loan rates with double-interaction terms

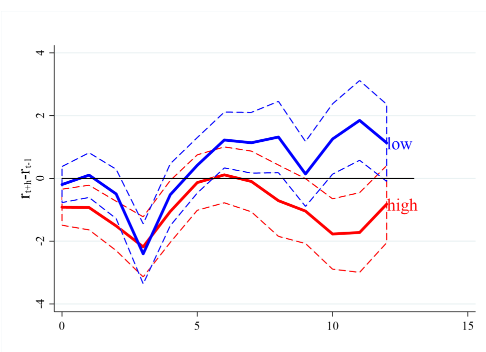
(a) Bank concentration, easing



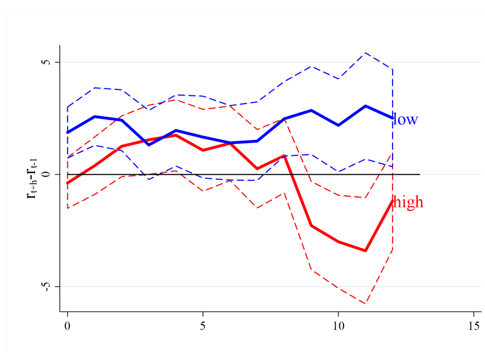
(b) Bank concentration, tightening



(c) Bank capital ratio, easing



(d) Bank capital ratio, tightening



Notes: Impulse response functions to a monetary shock during tightening and easing periods for a high and low level of bank concentration or capital ratio: $\beta^{+/-,h} + \gamma^{+/-,h}\,(m^{HHI,\%} \pm 2sd^{HHI,\%})$.

B Model Appendix

B.1 Full Model

Next to the afore-described financial intermediaries, the full model includes two types of households, entrepreneurs, labor packers and unions, capital and final goods producers, and a monetary authority following Gerali et al. (2010). The baseline environment deviates from Gerali et al. (2010) in two ways: the impatient household and entrepreneur do not face a credit constraint, and the only source of uncertainty is a monetary shock.

B.1.1 Patient and Impatient Households

There is a unit mass of patient and impatient households, each denoted by i. In the baseline model, both types of households differ only in terms of their subjective discount factor β^{χ}, with $\chi \in \{P, I\}$, where $\beta^P > \beta^I$.[46] Otherwise, the households preferences are the same. Both types consume, work, and own a housing stock, which is in aggregate in fixed supply.[47] Each household i of type $\chi \in \{P, I\}$ maximizes expected utility:

$$\mathbb{E}_t \sum_{t=0}^{\infty} \beta^{\chi,t} \left[(1 - a^{\chi}) \log \left(c_t^{\chi}(i) - a^{\chi} c_{t-1}^{\chi} \right) + \epsilon^h \log h_t^{\chi}(i) - \frac{l_t^{\chi}(i)^{1+\phi}}{1+\phi} \right],$$

depending on current consumption, $c_t^{\chi}(i)$, past aggregate consumption, c_{t-1}^{χ}, housing stock, $h_t^{\chi}(i)$, and, individual labor supplied, $l_t^{\chi}(i)$. a^{χ} governs the degree of external, group-specific habit formation.[48] ϕ measures disutility of labor. The utility of housing follows a log form governed by ϵ^h. The budget constraints differ across households, as the patient household provides deposits to the banking system, and the impatient household demands loans from the banking system.

The patient household's budget constraint follows:

$$c_t^P(i) + q_t^h \left(h_t^P(i) - h_{t-1}^P(i) \right) + d_t^P(i) \leq w_t^P l_t^P(i) + \left(1 + r_{t-1}^d \right) \frac{d_{t-1}^P(i)}{\pi_t} + \tau_t^P(i),$$

where $d_t^P(i)$ is patient household's deposit holding earning with gross interest income $1 + r_{t-1}^d d_{t-1}^P(i)/\pi_t$, w_t^P, real wage, q_t^h, price of housing, and $\tau_t^P(i)$ includes transfers from final goods producer and labor union, as these belong to the patient household.[49]

[46] The model extension with financial constraints adds a borrowing constraint to the impatient household.
[47] The housing market market-clearing condition is: $\bar{h} = h_t^P + h_t^I$, with constant housing supply, $\bar{h}$.
[48] Setting a^{χ} to 0 nests the case without habit. Multiplying by $(1 - a^{\chi})$ cancels out steady-state distortions.
[49] The bank does not pay a dividend and retains profits for next period's bank capital.

The impatient household's budget constraint follows:

$$c_t^I(i) + q_t^h\left(h_t^I(i) - h_{t-1}^I(i)\right) + b_{t-1}^I(i)\left(1 + r_{t-1}^{bH}\right)/\pi_t \leq w_t^I l_t^I(i) + b_t^I(i),$$

where $b_t^I(i)$ reflects impatient household's outstanding debt with gross interest expenses $1 + r_{t-1}^{bH} b_{t-1}^I(i)/\pi_t$, and, w_t^I, impatient household's real wage.

B.1.2 Entrepreneurs

A unit mass of entrepreneurs i produces a homogeneous intermediate good using two inputs: capital, k_t^E, purchased from capital-good producers, and hired labor input from the patient, l_t^P, and impatient household, l_t^I. Similar to the households, the entrepreneur's utility depends on current individual consumption, $c_t^E(i)$, and lagged aggregate consumption, c_{t-1}^E, governed by a^E. The entrepreneur maximizes expected utility:

$$\mathbb{E}_t \sum_{t=0}^{\infty} \beta_E^t \log\left(c_t^E(i) - a^E c_{t-1}^E\right),$$

subject to entrepreneur's budget constraint:

$$\begin{aligned} c_t^E(i) + w_t^I l_t^I(i) + w_t^P l_t^P(i) + \frac{1 + r_{t-1}^{bE}}{\pi_t} b_{t-1}^E(i) + q_t^k k_t^E(i) + \upsilon(u_t(i)) k_{t-1}^E(i) \leq \\ \frac{y_t^E(i)}{x_t} + b_t^E(i) + (1-\delta) q_t^k k_t^E(i), \end{aligned}$$

where $b_t^E(i)$ is the entrepreneur's outstanding debt with gross interest expenses $1 + r_{t-1}^{bE} b_{t-1}^E(i)/\pi_t$, q_t^k, the price of physical capital, δ, the depreciation rate, $\upsilon(u_t(i))$, capital utilization costs, $w_t^I l_t^I(i)$ and $w_t^P l_t^P(i)$, the wage bill for hiring labor from impatient and patient households, x_t, the price markup, and, $y_t^E(i)$, the produced wholesale good. The production function follows:

$$y_t^E(i) = \left[u_t(i) k_{t-1}^E(i)\right]^\alpha \left[l_t^E(i)\right]^{1-\alpha} = \left[u_t(i) k_{t-1}^E(i)\right]^\alpha \left[\left(l_t^P(i)\right)^\mu \left(l_t^I(i)\right)^{(1-\mu)}\right]^{1-\alpha}.$$

The labor input from the two types of households is combined to aggregate labor input, $l_t^E(i) = \left(l_t^P(i)\right)^\mu \left(l_t^I(i)\right)^{(1-\mu)}$, with μ governing the patient household's labor income share.

B.1.3 Labor Packers and Labor Unions

Perfectly competitive labor packers bundle differentiated labor inputs m using a CES aggregator and sell the homogenized bundle to the labor union. The labor union then provides the homogenized labor bundle to the entrepreneur as input. There exist two unions χ for each type of labor input m, with $\chi \in \{I, P\}$ for the impatient and patient household. Each labor union sets nominal wage, W_t^χ, subject to the entrepreneur's downward-sloping labor demand, and Rotemberg adjustment costs, κ_w. To cover for adjustment costs, the union charges a lump-sum fee and maximizes:

$$\mathbb{E}_t \sum_{t=0}^{\infty} \beta_u^t \Bigg\{ \Lambda_t^\chi(i,m) \left[\frac{W_t^\chi(m)}{P_t} l_t^\chi(i,m) - \frac{\kappa_w}{2} \left(\frac{W_t^\chi(m)}{W_{t-1}^\chi(m)} - \pi_{t-1}^{\iota_w} \pi^{1-\iota_w} \right)^2 \frac{W_t^\chi}{P_t} \right] - \frac{l_t^\chi(i,m)^{1+\phi}}{1+\phi} \Bigg\},$$

subject to labor demand $l_t^\chi(i,m) = \left(\frac{W_t^\chi(m)}{W_t^\chi}\right)^{-\epsilon^l} l_t^\chi$, where ϵ^l measures the degree substitutability. The labor union discounts future income with stochastic discount factor, $\Lambda_t^\chi(i,m)$, of the respective household. Adjustment costs incur relative to a weighted average of steady-state, $\pi^{1-\iota_w}$, and lagged inflation, $\pi_{t-1}^{\iota_w}$, with weight ι_w on lagged inflation.

In the symmetric equilibrium, labor supply of household with type χ is:

$$\kappa_w \left(\pi_t^{w,\chi} - \pi^{\iota_w, t-1} \pi^{1-\iota_w} \right) \pi_t^{w,\chi} = \beta^\chi \mathbb{E}_t \left[\frac{\Lambda_{t+1}^\chi}{\Lambda_t^\chi} \kappa_w \left(\pi_{t+1}^{w,\chi} - \pi_{t-1}^{\iota_w} \pi^{1-\iota_w} \right) \right] + \left(1 - \epsilon^l\right) l_t^\chi + \frac{\epsilon^l l_t^{\chi,1+\phi}}{w_t^\chi \Lambda_t^\chi},$$

where nominal wage inflation is defined as $\pi_t^{w,\chi} = \frac{W_t^\chi}{W_{t-1}^\chi}$ and the real wage as $w_t^\chi = \frac{W_t^\chi}{P_t}$.

B.1.4 Capital and Final Goods Producers

The capital good producer operates under perfect competition and purchases last period's depreciated physical capital stock, $\left(1 - \delta^k\right) k_{t-1}$, at a price q_t^k from the entrepreneur, and i_t units of the final good from retailers at a price P_t. The capital good producer converts the two input goods into new physical capital subject to quadratic investment adjustment costs, governed by cost parameter κ_i. It sells new capital back to entrepreneurs at the same price q_t^k. The capital good producers objective is to maximize the sum of expected future profits discounted by the entrepreneur's stochastic discount factor, $\Lambda^E_{0,t}$:

$$\mathbb{E}_t \sum_{t=0}^{\infty} \Lambda^E_{0,t} \left(q^k_t \left[k_t - (1-\delta^k) k_t \right] - i_t \right)$$

subject to the evolution of capital:

$$k_t = (1-\delta^k) k_{t-1} + \left[1 - \frac{\kappa_i}{2} \left(\frac{i_t}{i_{t-1}} - 1 \right)^2 \right] i_t.$$

The final good firms operate under monopolistic competition. Each final good firm j buys intermediate goods from entrepreneurs at wholesale price, P^W_t, differentiates goods at no cost, and sells them to customers as a final good. Retail prices are sticky and indexed to an average of past and steady-state price inflation with weight ι_p on past inflation. The firm incurs Rotemberg adjustment costs, κ_p, for changing prices beyond indexation. The final price, $P_t(j)$, is chosen to maximize profits:

$$\mathbb{E}_t \sum_{t=0}^{\infty} \Lambda^P_{0,t} \left[P_t(j) y_t(j) - P^W_t y_t(j) - \frac{\kappa_p}{2} \left(\frac{P_t(j)}{P_{t-1}} - \pi^{\iota_p}_{t-1} \pi^{1-\iota_p} \right)^2 P_t y_t \right],$$

subject to final good demand of good j with demand price elasticity ε^y:

$$y_t(j) = \left(\frac{P_t(j)}{P_t} \right)^{-\varepsilon^y_t} y_t.$$

B.1.5 Monetary Policy and Market Clearing

The central bank follows a standard Taylor rule:

$$(1+r^f_t) = (1+r^f)^{(1-\phi_R)} (1+r^f_{t-1})^{\phi_R} \left(\frac{\pi_t}{\pi} \right)^{\phi_\pi \, (1-\phi_R)} \left(\frac{y_t}{y_{t-1}} \right)^{\phi_y \, (1-\phi_R)} \varepsilon^R_t,$$

where ϕ_R reflects the weight on the lagged policy rate, ϕ_π and ϕ_y, the responsiveness to inflation and output growth, and ε^R_t an i.i.d. monetary shock with standard deviation σ_R.

The goods market market-clearing condition is:

$$y_t = c^E_t + c^P_t + c^I_t + q^k_t \left[k_t - (1-\delta) k_{t-1} \right] + k_{t-1} \phi(u_t) + \delta^{KB} \frac{K^{KB}_{t-1}}{\pi_t} + Adj_t.$$

where Adj_t combines all adjustment costs (prices, wages, and banks).

B.2 Equilibrium Equations

$$c_t^I + q_t^h \left(h_t^I - h_{t-1}^I\right) + \left(1 + r_{t-1}^{BH}\right) \frac{b_{t-1}^I}{\pi_t} = w_t^I l_t^I + b_t^I \tag{25}$$

$$\frac{\left(1 - a^I\right)}{c_t^I - a^I c_{t-1}^I} = \lambda_t^I \tag{26}$$

$$\lambda_t^I q_t^h = \frac{\varepsilon^h}{h_t^I} + \beta^I \mathbb{E}_t \left[\lambda_{t+1}^I q_{t+1}^h\right] \tag{27}$$

$$\lambda_t^I = \beta^I \mathbb{E}_t \lambda_{t+1}^I \frac{\left(1 + r_t^{BH}\right)}{\pi_{t+1}} \tag{28}$$

$$\kappa_w \left(\pi_t^{w,I} - \pi_{t-1}^{\iota_w} \pi^{1-\iota_w}\right) \pi_t^{w,I} = \beta^I \mathbb{E}_t \frac{\lambda_{t+1}}{\lambda_t} \kappa_w \left(\pi_{t+1}^{w,I} - \pi_t^{\iota_w} \pi^{1-\iota}\right) \frac{\left(\pi_{t+1}^{w,I}\right)^2}{\pi_{t+1}} + \left(1 - \varepsilon^l\right) l_t^I + \frac{\varepsilon^l \left(l_t^I\right)^{1+\phi}}{w_t^{w,I} \lambda_t^I} \tag{29}$$

$$\pi_t^{w,I} = \frac{w_t^{w,I}}{w_{t-1}^{w,I}} \pi_t \tag{30}$$

$$c_t^P + q_t^h \left(h_t^P - h_{t-1}^P\right) + d_t^P = w_t^P l_t^P + \left(1 + r_{t-1}^d\right) \frac{d_{t-1}^P}{\pi_t} + \tau_t^P \tag{31}$$

$$\frac{\left(1 - a^P\right)}{c_t^P - a^P c_{t-1}^P} = \lambda_t^P \tag{32}$$

$$\lambda_t^P q_t^h = \frac{\varepsilon^h}{h_t^P} + \beta^P \mathbb{E}_t \lambda_{t+1}^P q_{t+1}^h \tag{33}$$

$$\lambda_t^P = \beta^P \mathbb{E}_t \lambda_{t+1}^P \frac{\left(1 + r_t^d\right)}{\pi_{t+1}} \tag{34}$$

$$\kappa_w \left(\pi_t^{w,P} - \pi_{t-1}^{\iota_w} \pi^{1-\iota_w}\right) \pi_t^{w,P} = \beta^P \mathbb{E}_t \frac{\lambda_{t+1}}{\lambda_t} \kappa_w \left(\pi_{t+1}^{w,P} - \pi_t^{\iota_w} \pi^{1-\iota}\right) \frac{\left(\pi_{t+1}^{w,P}\right)^2}{\pi_{t+1}} + \left(1 - \varepsilon^l\right) l_t^P + \frac{\varepsilon^l \left(l_t^P\right)^{1+\phi}}{w_t^{w,P} \lambda_t^P} \tag{35}$$

$$\pi_t^{w,P} = \frac{w_t^{w,P}}{w_{t-1}^{w,P}} \pi_t \tag{36}$$

$$c_t^E + w_t^P l_t^P + w_t^I l_t^I + \left(1 + r_{t-1}^{bE}\right) b_{t-1}^E / \pi_t + q_t^k k_t^E + \upsilon(u_t) k_{t-1}^E = \frac{y_t^E}{x_t} + b_t^E + q_t^k \left(1 - \delta\right) k_{t-1}^E \tag{37}$$

$$\upsilon(u_t) = \zeta_1 (u_t - 1) + \zeta_2 (u_t - 1)^2 \tag{38}$$

$$r_t^k = \zeta_1 + \zeta_2 (u_t - 1) \tag{39}$$

$$\frac{\left(1 - a^E\right)}{c_t^E - a^E c_{t-1}^E} = \lambda_t^E \tag{40}$$

$$\lambda_t^E = \beta^E \mathbb{E}_t \left[\lambda_{t+1}^E \frac{\left(1 + r_t^{bE}\right)}{\pi_{t+1}}\right] \tag{41}$$

$$\lambda_t^E q_t^k = \beta^E \mathbb{E}_t \left\{\lambda_{t+1}^E \left[r_{t+1}^k u_{t+1} + q_{t+1}^k (1 - \delta) - \left(\zeta_1 (u_{t+1} - 1) + \frac{\zeta_2}{2} (u_{t+1} - 1)^2\right)\right]\right\} \tag{42}$$

$$y_t^E = \left[u_t k_{t-1}^E\right]^\alpha \left[\left(l_t^P\right)^\mu \left(l_t^I\right)^{(1-\mu)}\right]^{1-\alpha} \tag{43}$$

$$w_t^P = \mu (1 - \alpha) \frac{y_t^E}{l_t^P} \frac{1}{x_t} \tag{44}$$

$$w_t^I = (1 - \mu)(1 - \alpha) \frac{y_t^E}{l_t^I} \frac{1}{x_t} \tag{45}$$

$$r_t^k = \alpha \left[u_t k_{t-1}^E\right]^{\alpha - 1} \left[\left(l_t^P\right)^\mu \left(l_t^I\right)^{(1-\mu)}\right]^{1-\alpha} \tag{46}$$

$$k_t = (1 - \delta) k_{t-1} + \left[1 - \frac{\kappa_i}{2} \left(\frac{i_t}{i_{t-1}} - 1\right)^2\right] i_t \tag{47}$$

$$1 = q_t^k \left[1 - \frac{\kappa_i}{2} \left(\frac{i_t}{i_{t-1}} - 1\right)^2 - \kappa_i \left(\frac{i_t}{i_{t-1}} - 1\right) \frac{i_t}{i_{t-1}}\right] + \beta^E \mathbb{E}_t \frac{\lambda_{t+1}^E}{\lambda_t^E} q_{t+1}^k \kappa_i \left(\frac{i_{t+1}}{i_t} - 1\right) \left(\frac{i_{t+1}}{i_t}\right)^2 \tag{48}$$

$$\Pi_t^r = y_t \left(1 - \frac{1}{x_t}\right) - \frac{\kappa_p}{2} \left(\pi_t - \pi_{t-1}^{\iota_p} \pi^{1-\iota_p}\right)^2 \tag{49}$$

$$0 = 1 - \varepsilon^y + \frac{\varepsilon^y}{x_t} - \kappa_p \left(\pi_t - \pi_{t-1}^{\iota_p} \pi^{1-\iota_p}\right) \pi_t + \beta^P \mathbb{E}_t \left[\frac{\lambda_{t+1}^P}{\lambda_t^P} \kappa_p \left(\pi_{t+1} - \pi_t^{\iota_p} \pi^{1-\iota_p}\right) \pi_{t+1} \frac{y_{t+1}}{y_t}\right] \tag{50}$$

$$B_t = d_t^P + K_t^b \tag{51}$$

$$\pi_t K_t^b = \left(1 - \delta^b\right) K_{t-1} + \Pi_{t-1}^b \tag{52}$$

$$\left(R_t^b - r_t^f\right) = -\kappa_{KB}\left(\frac{K_t^b}{B_t} - \nu^b\right)\left(\frac{K_t^b}{B_t}\right)^2 \tag{53}$$

$$\Pi_t^b = r_t^{bH} b_t^{bH} + r_t^{bE} b_t^{bE} - r_t^d d_t^P - \frac{\kappa_{KB}}{2}\left(\frac{K_t^b}{B_t} - \nu^b\right)^2 K_t^b - \frac{\kappa_d}{2}\left(\frac{d_t^P}{d_{ss}^P} - 1\right)^2 r_t^d d_t^P -$$
$$\frac{\kappa_{bH}}{2}\left(\frac{b_t^{bH}}{b_{ss}^{bH}} - 1\right)^2 r_t^{bH} b_t^{bH} - \frac{1}{\psi_{bH}^2}\left\{\exp\left[\psi_{bH}\left(\frac{b_t^{bH}}{b_{ss}^{bH}} - 1\right)\right] - \psi_{bH}\left(\frac{b_t^{bH}}{b_{ss}^{bH}} - 1\right) - 1\right\} r_t^{bH} b_t^{bH}$$
$$\frac{\kappa_{bE}}{2}\left(\frac{b_t^{bE}}{b_{ss}^{bE}} - 1\right)^2 r_t^{bH} b_t^{bE} - \frac{1}{\psi_{bE}^2}\left\{\exp\left[\psi_{bE}\left(\frac{b_t^{bE}}{b_{ss}^{bE}} - 1\right)\right] - \psi_{bE}\left(\frac{b_t^{bE}}{b_{ss}^{bE}} - 1\right) - 1\right\} r_t^{bE} b_t^{bE} \tag{54}$$

$$\left(\epsilon^d - 1\right) - \epsilon^d \frac{r_t^f}{r_t^d} + \epsilon^d \kappa_d \left(\frac{d_t^P}{d_{ss}^P} - 1\right)\frac{d_t^P}{d_{ss}^P} = 0 \tag{55}$$

$$-\left(\epsilon^{bH} - 1\right) + \frac{\epsilon^{bH} R_t^b}{r_t^{bH}} + \epsilon^{bH}\kappa_{bH}\left(\frac{b_t^{bH}}{b_{ss}^{bH}} - 1\right)\frac{b_t^{bH}}{b_{ss}^{bH}} + \frac{\epsilon^{bH}}{\psi_{bH}}\left\{\exp\left[\psi_{bH}\left(\frac{b_t^{H}}{b_{ss}^{bH}} - 1\right)\right] - 1\right\}\frac{b_t^{bH}}{b_{ss}^{H}} = 0 \tag{56}$$

$$-\left(\epsilon^{bE} - 1\right) + \frac{\epsilon^{bE} R_t^b}{r_t^{bE}} + \epsilon^{bE}\kappa_{bE}\left(\frac{b_t^{bE}}{b_{ss}^{bE}} - 1\right)\frac{b_t^{bE}}{b_{ss}^{bE}} + \frac{\epsilon^{bE}}{\psi_{bE}}\left\{\exp\left[\psi_{bE}\left(\frac{b_t^{bE}}{b_{ss}^{bE}} - 1\right)\right] - 1\right\}\frac{b_t^{bE}}{b_{ss}^{bE}} = 0 \tag{57}$$

$$(1 + r_t^f) = (1 + r^f)^{(1-\phi_R)}(1 + r_{t-1}^f)^{\phi_R}\left(\frac{\pi_t}{\pi}\right)^{\phi_\pi\ (1-\phi_R)}\left(\frac{y_t}{y_{t-1}}\right)^{\phi_y\ (1-\phi_R)} \varepsilon_t^R \tag{58}$$

$$y_t = c_t^E + c_t^P + c_t^I + q_t^k\left[k_t - (1-\delta)\,k_{t-1}\right] + k_{t-1}\phi(u_t) + \delta^{KB}\frac{K_{t-1}^{KB}}{\pi_t} + Adj_t \tag{59}$$

$$\bar{h} = h_t^P + h_t^I \tag{60}$$

$$B_t = b_t^{bH} + b_t^{bE} \tag{61}$$

$$Y_t = c_t^E + c_t^P + c_t^I + i_t \tag{62}$$

B.3 Calibration of Baseline Model

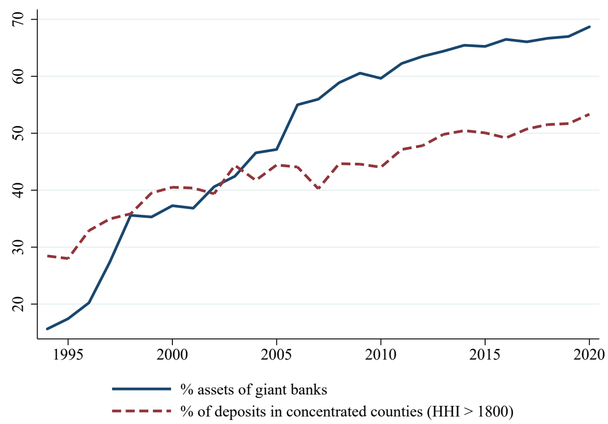
Table B.1: Calibration of model parameters following Gerali et al. (2010)

Parameter	**Description**	**Value**
κ^{Kb}	Adjustment costs of bank capital ratio	11.49
δ^b	Management cost of bank	0.1049[a]
β^P	Discount factor of patient household	0.9943
$\beta^{I,E}$	Discount factor of impatient household and entrepreneur	0.975[b]
ϕ	Inverse of Frisch elasticity of labor supply	1
ϵ^h	Housing preference	0.2
$a^{P,I,E}$	Habit consumption	0.86
$\epsilon^{m,I}$	Steady-state LTV-ratio for impatient households	0.7[c]
α	Output elasticity with respect to capital	0.25
μ	Labor cost share of patient households costs	0.8
ζ_1	Adjustment costs for capacity utilization	0.0478
ζ_2	Adjustment costs for capacity utilization	0.00478
$\epsilon^{m,E}$	Steady-state LTV-ratio for entrepreneur	0.35[c]
κ_w	Adjustment costs of wages	99.9
ι_w	Indexation of wage inflation to past wage inflation	0.28
ϵ^l	Steady-state labor market markup	5
δ	Depreciation rate of physical capital	0.025
κ_i	Adjustment costs of investment	10.18
κ_p	Adjustment costs of good prices	28.65
ι_p	Indexation of price inflation to past price inflation	0.16
ϵ^y	Steady-state goods market markup	6
ϕ_R	Taylor rule smoothing parameter	0.77
ϕ_π	Taylor rule response to inflation	1.98[d]
ϕ_x	Taylor rule response to output	0.35
σ_r	Standard deviation of monetary shock	0.002

[a] δ^b varies with $\epsilon^d, \epsilon^{bh}, \epsilon^{be}, \nu^b$ to satisfy in the steady state $\delta^b = \Pi^b / K^b$.
[b] In the baseline model without borrowing constraints $\beta^{I,E}$ depends on $\beta^P, \epsilon^d, \epsilon^{bH}, \epsilon^{bE}$.
[c] Only used in the model with borrowing constraints.
[d] In Section 6.3, the coefficient on inflation is higher (2.9) to avoid indeterminacy issues.

B.4 Calibration of Heterogeneous Bank Model

Figure B.1: Share of high-concentration markets and giant banks over time



Notes: The deposit-weighted market share of high-concentration markets from 1994 to 2019. The cutoff for high-concentration counties is 1800, following the Department of Justice 's classification defining markets with an HHI above 1800 points as highly concentrated. The share of assets held by banks with more than $100 billion in assets (in $2018). Source: Federal Deposit Insurance Corporation, Department of Justice.

B.5 Micro-founding Asymmetries

While most of the literature focused on price adjustment costs in banking to explain an incomplete monetary pass-through (Levieuge and Sahuc, 2021), my paper argues that quantity adjustment costs are more in line with anecdotal evidence. The motivation is that banks effectively incur charges of expanding lending (e.g., additional overhead, screening costs). In summer 2019, newspapers reported that banks struggled to meet the abnormally high demand for refinancing as 30-year mortgage rates declined. Consequently, the days to close a purchase loan increased to 60 days, typically averaging 40 days. At some of the Wells Fargo locations, it took more than 120 days to close, according to Mortgage Professional America.[50] Similarly, a Bloomberg article argues that banks could have offered lower rates if they would have hired more operational personnel.[51]

To incorporate the evidence for higher costs to scale up lending, I assume that adjustment costs are asymmetric and incur in terms of percent deviations from steady-state lending $\left(\frac{B_t}{B_{ss}} - 1\right)$. A convenient modeling approach uses an altered linex (linear, exponential) cost function.[52] The advantage of an altered linex function is that the function is still continuous, differentiable, and the model can be solved with perturbation methods. As shown in equation (63), the function consists of a quadratic and an asymmetric part. The asymmetric part builds on an exponential function, converging to zero when the argument declines. κ_l measures the cost function's concavity and ψ_l the degree of asymmetry. The altered linex function nests the symmetric case when ψ_l approaches zero.

$$C\left(\frac{B_t}{B_{ss}} - 1\right) = \underbrace{\frac{\kappa_l}{2}\left(\frac{B_t}{B_{ss}} - 1\right)^2}_{\text{symmetric part}} + \underbrace{\frac{1}{\psi_l^2}\left\{\exp\left[\psi_l\left(\frac{B_t}{B_{ss}} - 1\right)\right] - \psi_l\left(\frac{B_t}{B_{ss}} - 1\right) - 1\right\}}_{\text{asymmetric ("linex") part}} \tag{63}$$

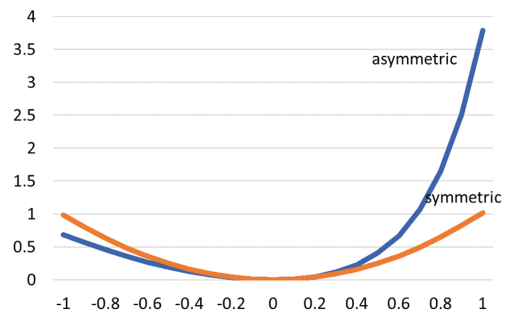
Figure B.2 presents the cost function in terms of deviations from steady-state lending. Loan expansions are located on the right of 0 with adjustment costs exponentially increasing for $\psi_l > 0$.

[50] Mortgage Professional America's article on "Big banks fumble the ball on spiking mortgage demand" (link).

[51] Bloomberg's article on "Banks fire up their mortgage machine for a new refinancing boom" (link).

[52] See for example, Levieuge and Sahuc (2021) for modeling downward loan rate rigidity or Abbritti and Fahr (2013) and Fahr and Smets (2010) for downward wage rigidity.

Figure B.2: Altered linex cost function



Notes: The altered linex cost function is shown for two different parameterizations. In the symmetric case, ψ_l is 10, and κ_l is 1. In the asymmetric case, ψ_l is 50, and κ_l is 1, which shifts the costs up in the positive range.

An alternative way to generate asymmetric monetary policy pass-through is related to bank capital requirements. Building on the intuition that banks face a minimum bank capital ratio and hence undershooting the bank capital ratio is much more costly than overshooting, assume asymmetric bank capital adjustment costs at the headquarters level. The adjustment costs, $\mathbb{A}_{KB}$, in terms of deviations from the capital ratio, $\left(\frac{K_t^b}{B_t} - \nu^b\right)$, take the following form:

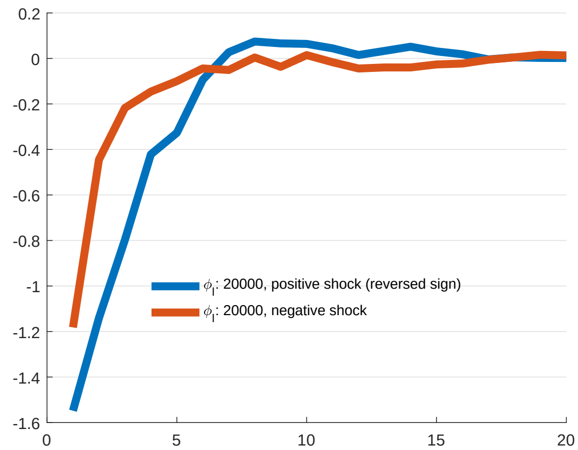
$$\mathbb{A}_{KB}\left(\frac{K_t^b}{B_t}\right) = \frac{\kappa_{KB}}{2}\left(\frac{K_t^b}{B_t} - \nu^b\right)^2 + \frac{1}{\psi_{KB}^2}\left\{\exp\left[-\psi_{KB}\left(\frac{K_t^b}{B_t} - \nu^b\right)\right] + \psi_{KB}\left(\frac{K_t^b}{B_t} - \nu^b\right) - 1\right\}$$

Note that in this case, the argument in parentheses is negative as there are higher costs for low levels of the bank capital ratio. The wholesale funding rate in terms of the bank capital ratio and the policy rate, r_t^f, turns to:

$$R_t^b = r_t^f - \kappa_{KB}\left(\frac{K_t^b}{B_t} - \nu^b\right)\left(\frac{K_t^b}{B_t}\right)^2 - \frac{1}{\psi_{KB}}\left\{1 - \exp\left[-\psi_{KB}\left(\frac{K_t^b}{B_t} - \nu^b\right)\right]\right\}\left(\frac{K_t^b}{B_t}\right)^2$$

Figure B.3 presents the loan rate pass-through to monetary policy shock for monetary tightening and easing in the presence of adjustment coasts at the bank head quarter. The pass-through is much smaller in times of monetary easing (negative shock).

Figure B.3: Impulse response of the loan rate to monetary tightening vs. easing



Notes: Impulse response functions to a negative and positive (reversed sign) monetary shock with asymmetric adjustment costs at the bank headquarters for low levels of the bank capital ratio.

B.6 Extension Heterogeneous Branches

This section presents the details of the heterogeneous bank branch extension and the derivation of the aggregate markup. To keep the framework tractable, I consider two spatially segmented types of the retail loan and deposit branches differing in the elasticity of loan demand and deposit supply.

Deposit branch types. In two spatially segmented markets, a continuum of retail deposit branches collects deposits from customers. The deposit branches transfer the deposits to the wholesale unit and earn a positive spread due to location-specific monopolistic competition in location c. Each branch maximizes profits by choosing a location-specific deposit rate denoted with the superscript c, $r_t^{d,c}$, taking into account the wholesale lending rate, R_t^d, and the location-specific deposit supply from the households, $d^p(r_t^{d,c}) = \left(\frac{r_t^{d,c}}{\bar{r}_t^{d,c}}\right)^{-\epsilon^{d,c}} \bar{d}_t^{p,c}$:

$$\max_{r_t^{d,c}} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[R_t^{d,c} d^p(r_t^{d,c}) - r_t^{d,c} d^p(r_t^{d,c}) - \mathbb{A}_D \left(d^p(r_t^{d,c}) \right) \bar{r}_t^{d,c} \bar{d}_t^{p,c} \right], \tag{64}$$

where $\bar{r}_t^{d,c}$ and $\bar{d}_t^{p,c}$ reflect the aggregate deposit rate and the aggregate deposit supply in location c. $\mathbb{A}_D \left(\frac{d_t^{p,c}}{d_{ss}^{p,c}} \right)$ are quadratic adjustment costs in terms of deviations from the steady-state level of branch deposits $d_{ss}^{p,c}$. The retail deposit branch's deposit rate condition is:

$$\left(\epsilon^{d,c} - 1\right) - \epsilon^{d,c} \frac{R_t^d}{r_t^{d,c}} + \epsilon^{d,c} \kappa_d \left(\frac{d_t^{p,c}}{d_{ss}^{p,c}} - 1 \right) \frac{d_t^{p,c}}{d_{ss}^{p,c}} = 0 \tag{65}$$

Loan branch types. In two spatially segmented markets, a continuum of loan branches finances loans to households (bH) and entrepreneurs (bE). The loan branches obtain funding from the wholesale unit at the wholesale funding rate, R_t^b, and earn a positive spread due to location-specific monopolistic competition in location c. The loan branches maximize profits by choosing the location-specific loan rate taking as given the wholesale funding rate, R_t^b, and a location-specific loan demand: $b_t^{l,c}(r_t^{l,c}) = \left(\frac{r_t^{l,c}}{\bar{r}_t^{l,c}}\right)^{-\epsilon^{l,c}} \bar{b}_t^{l,c}$ with $l \in \{bE, bH\}$. The branch for l loans in location c solves:

$$\max_{r_t^{l,c}} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[r_t^{l,c} b_t^{l,c}(r_t^{l,c}) - R_t^b b_t^{l,c}(r_t^{l,c}) - \mathbb{A}_b \left(b_t^{l,c}(r_t^{l,c}) \right) \bar{r}_t^{l,c} \bar{b}_t^{l,c} \right], \tag{66}$$

where $\bar{r}_t^{l,c}$ and $\bar{b}_t^{l,c}$ reflect the aggregate loan rate and loan demand. $\mathbb{A}_b$ are quadratic adjustment costs in terms of deviations from the steady-state level of l loans in location c.

The loan branch's optimality condition is:

$$-\left(\epsilon^{l,c}-1\right)+\epsilon^{l,c}\frac{R_t^b}{r_t^{l,c}}+\epsilon^{l,c}\kappa_l\left(\frac{b_t^{l,c}}{b_{ss}^{l,c}}-1\right)\frac{b_t^{l,c}}{b_{ss}^{l,c}}=0 \tag{67}$$

Wholesale unit. The wholesale unit manages the flow of funds between the deposit and loan branches in low- and high-concentration markets denoted by the superscripts l and h. The wholesale unit's balance sheet becomes: $b_t^{bH,l}+b_t^{bH,h}+b_t^{bE,l}+b_t^{bE,h}=d_t^{p,l}+d_t^{p,h}+K_t^b$.

Equilibrium. To close the model, assume an exogenous market size distribution with a share α^m of ϵ^l regions. This implies the following loan and deposit relationships, $b_t^{h,l}=\frac{\alpha^m}{1-\alpha^m}b_t^{h,h}$ and $d_t^{p,l}=\frac{\alpha^m}{1-\alpha^m}d_t^{p,h}$. The aggregate loan and deposit rates are: $r_t^{bH}=\alpha^m r^{bH,l}+(1-\alpha^m)r^{bH,h}$, $r_t^{bE}=\alpha^m r^{bE,l}+(1-\alpha^m)r^{bE,h}$ and $r_t^d=\alpha^m r^{d,l}+(1-\alpha^m)r^{d,h}$.

Analytical expression for the "aggregate markup." I derive an analytical expression for the "aggregate markup," a sufficient statistic summarizing the degree of heterogeneity between branches and market shares that also informs about the monetary policy pass-through. I further evaluate the impact of changes in the *extensive* margin, α^m, and the *intensive* margin, ϵ^l.

Proof: Proposition 1. Since the household's saving and investment decision concerns aggregate rates, the aggregate rate, $\bar{r}_t$, equals the weighted sum of rates r_t^l in low- and in r_t^h high-concentration markets governed by α^m, the market share of the low-concentration region:

$$\bar{r}_t=\alpha^m r_t^l+(1-\alpha^m)\,r_t^h \tag{68}$$

After substituting in the two branch type rate setting functions of the branches from equations (64) and (67), dividing through R_t^b and abstracting from adjustment costs for tractability, the aggregate markup $\bar{m}$ simplifies to a function of exogenous parameters α^m, ϵ^l and ϵ^h only:

$$\bar{m}=\alpha^m\underbrace{\left(\frac{\epsilon^l}{\epsilon^l-1}\right)}_{m^l}+(1-\alpha^m)\underbrace{\left(\frac{\epsilon^h}{\epsilon^h-1}\right)}_{m^h} \tag{69}$$

□

Proof: Proposition 2. Based on the empirical evidence on loan rates, recall that $|\epsilon^l|>|\epsilon^h|$ and $m^l<m^h$. First, examine the partial effect of changes in the extensive margin α^m on the

aggregate markup:

$$\frac{\partial \bar{m}}{\partial \alpha^m} = \left(\frac{\epsilon^l}{\epsilon^l - 1}\right) - \left(\frac{\epsilon^h}{\epsilon^h - 1}\right) < 0. \tag{70}$$

The aggregate markup decreases in the share of low-concentration markets. The sensitivity depends on the degree of heterogeneity, i.e., the difference between ϵ^l and ϵ^h. The more heterogeneous both markets, the larger the impact of changes in α^m. Turning the focus next to the partial effect of changes in the intensive margin, ϵ^h, shows that the markup decreases in the elasticity of loan demand in the high-concentration market, ϵ^h, and similarly depends on the relative share of the low-concentration market, α^m:

$$\frac{\partial \bar{m}}{\partial \epsilon^h} = (1 - \alpha^m)\left(\frac{\left(\epsilon^h - 1\right) - \epsilon^h}{\left(\epsilon^h - 1\right)^2}\right) = (1 - \alpha^m)\left(\frac{-1}{\left(\epsilon^h - 1\right)^2}\right) < 0. \tag{71}$$

□

B.7 Extension Heterogeneous Bank Headquarters

This section presents the details of the heterogeneous bank headquarters extension and comparative statics of the bank size distribution on marginal costs. To keep the structure tractable, I consider two heterogeneous bank types differing in the bank capital ratio.

Wholesale unit. There are two types of wholesale units j, corresponding to the headquarters of the regional and giant banks denoted with the superscript r and g. The wholesale units differ in terms of the bank capital requirement, $\nu^{b,j}\ \forall j \in \{r, g\}$. Each wholesale unit of type j maximizes profits from intermediating funds between loan, $b_t^{bH,j}$ and $b_t^{bE,j}$, and deposit branches, $d_t^{p,j}$, subject to the balance sheet constraint, $B_t^j = K_t^{b,j} + d_t^{p,j}$, and adjustment costs, $\mathbb{A}_{KB}$, in terms of the bank capital ratio, $\left(\frac{K_t^{b,j}}{B_t^j}\right)$:[53]

$$\max_{B_t^j, d_t^{p,j}} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[R_t^{b,j} B_t^j - R_t^d d_t^{p,j} - \mathbb{A}_{KB}\left(\frac{K_t^{b,j}}{B_t^j}\right) K_t^{b,j} \right]\ \forall\ j = \{r, g\}. \tag{72}$$

As wholesale unit's optimality condition in equation (73) shows, the wholesale funding rate of bank j depends on the bank capital ratio, $\nu^{b,j}$. Recall from Section 6.2.2 that in response to a positive monetary shock, the term in parentheses is positive. Hence, the

[53]The adjustment cost function equals: $\frac{\kappa_{KB}}{2}\left(\frac{K_t^{b,j}}{B_t^j} - \nu^{b,j}\right)^2$, where $B_t^j = b_t^{bH,j} + b_t^{bE,j}\ \forall\ j \in \{r, g\}$.

higher $\nu^{b,j}$, the less $R^{b,j}$ reacts.

$$R_t^{b,j} = R_t^d - \kappa_{KB}\left(\frac{K_t^{b,j}}{B_t^j} - \nu^{b,j}\right)\left(\frac{K_t^{b,j}}{B_t^j}\right)^2 \;\; \forall \;\; j = \{r, g\}. \tag{73}$$

Deposit branch types. There is a continuum of deposit branches belonging to each headquarters j. The deposit branches collect deposits from customers and stores these at wholesale unit j, earning a positive deposit spread due to monopolistic competition in the deposit market. Each branch maximizes profits by choosing the deposit rate taking as given the *uniform* wholesale lending rate, R^d, and deposit supply function, $d^p(r_t^d) = \left(\frac{r_t^d}{\bar{r}_t^d}\right)^{-\epsilon^d}\bar{d}_t^p$, as given:[54]

$$\max_{r_t^{d,j}} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[R_t^d d^p(r_t^{d,j}) - r_t^{d,j} d^p(r_t^{d,j}) - \mathbb{A}_D\left(d^p(r_t^{d,j})\right)\bar{r}_t^d \bar{d}^p\right] \tag{74}$$

where $\bar{r}_t^d$ and $\bar{d}_t^p$ reflect the aggregate deposit rate and aggregate deposits. $\mathbb{A}_D\left(\frac{d_t^{p,j}}{d_{ss}^{p,j}}\right)$ are quadratic adjustment costs in terms of deviations from steady state and relative to deposit expenses.[55] The deposit branch's optimality condition is:

$$-\epsilon^d \frac{R_t^d}{r_t^{d,j}} + \left(\epsilon^d - 1\right) + \epsilon^d \kappa_d \left(\frac{d_t^{p,j}}{d_{ss}^{p,j}} - 1\right)\frac{d_t^{p,j}}{d_{ss}^{p,j}} = 0 \tag{75}$$

Loan branches. There is a continuum of loan branches belonging to each headquarters j. The loan branches finance loans to households, b_t^{bH}, and to entrepreneurs, b_t^{bE}, with funding from the wholesale unit j at the headquarter-specific wholesale funding rate, $R_t^{b,j}$. The branches earn a positive loan spread due to monopolistic competition on the loan market and incur quadratic adjustment costs on adjusting lending. The loan branches maximize profits choosing the branch-specific loan rate taking as given the headquarter-specific wholesale funding rate $R_t^{b,j}$ and loan demand: $b_t^l(r_t^{l,j}) = \left(\frac{r_t^{l,j}}{\bar{r}_t^{l,}}\right)^{-\epsilon^l}\bar{b}_t^l$, $\forall\, l \in \{bE, bH\}$:

$$\max_{r_t^{l,j}} \mathbb{E}_t \sum_{t=0}^{\infty} \Lambda_{0,t}^P \left[r_t^{l,j} b_t^{l,j}(r_t^{l,j}) - R_t^{b,j} b_t^{l,j}(r_t^{l,j}) - \mathbb{A}_l\left(b_t^{l,j}(r_t^{l,j})\right)\bar{r}_t^l \bar{b}_t^{l,j}\right], \tag{76}$$

[54]The wholesale lending rate equals the policy rate in equilibrium and hence is the same across banks institutions.

[55]The adjustment costs take the following form, $\frac{\kappa_d}{2}\left(\frac{d^{p,j}(r_t^d)}{d_{ss}^{p,j}} - 1\right)^2$.

where $\bar{r}_t^l$ and $\bar{b}_t^l$ reflect the aggregate loan rate and aggregate loans. $\mathbb{A}_l$ are quadratic adjustment costs in terms of deviations from steady state.[56] The retail loan branch's optimality condition is:

$$-\left(\epsilon^l - 1\right) + \epsilon^l \frac{R_t^{b,j}}{r_t^{l,j}} + \epsilon^l \kappa_l \left(\frac{b_t^{l,j}}{b_{ss}^{l,j}} - 1\right) \frac{b_t^{l,j}}{b_{ss}^{l,j}} = 0 \tag{77}$$

Equilibrium. In equilibrium, aggregate loan supply and deposit demand across all banks equal loans demanded and deposits supplied by the households and firms. $B_t = B_t^r + B_t^g, b_t^{bH} = b_t^{bH,r} + b_t^{bH,g}, b_t^{bE} = b_t^{bE,r} + b_t^{bE,g}, d_t^p = d_t^{p,r} + d_t^{p,g}$. The aggregate rates are: $r_t^j = \alpha^b \left(r_t^{j,r}\right) + (1-\alpha^b) r_t^{j,g} \ \forall \ j \in \{d, bH, bE\}$, where α^b is the regional bank's market share. This implies for the bank capital, deposit and loan distributions: $K_t^{b,r} = \frac{1-\alpha^b}{\alpha^b} K_t^{b,g}$, $b_t^{bH,r} = \frac{1-\alpha^B}{\alpha^B} b_t^{bH,g}$, $b_t^{bE,r} = \frac{1-\alpha^B}{\alpha^B} b_t^{bE,g}$, $d_t^{p,r} = \frac{1-\alpha^d}{\alpha^d} d_t^{p,g}$.[57]

Analytics on the bank size distribution. How do compositional changes affect the wholesale funding rate, the bank's marginal cost for loans, and what is the impact of changes in the *extensive*, α^b, and *intensive*, ν^b, margin on the aggregate loan rate?

Proof: Proposition 3. Assume borrowers respond to an aggregate bundle of loans supplied by all banks. The aggregate loan rate, $\bar{r}_t^l$, is the weighted sum of rates by regional, r_t^r, and giant banks, r_t^g, governed by regional banks' market share, α^b:

$$\bar{r}_t^l = \alpha^b r_t^{l,r} + \left(1 - \alpha^b\right) r_t^{l,g}. \tag{78}$$

After substituting in the regional and gaint bank's loan rate decisions from equations (73) and (77), assuming equal market power across banks, and abstracting from adjustment costs, the aggregate loan rate, $\bar{r}_t^l$, simplifies to a function of parameters α^b, $\nu^{b,r}$ and $\nu^{b,g}$, the capital ratios, and the policy rate, r_t^f:

$$\bar{r}_t^l = \frac{\epsilon^l}{(\epsilon^l - 1)} \left(r_t^f - \alpha^b \kappa_{KB} \left(\frac{K_t^{b,r}}{B_t^r} - \nu^{b,r}\right) \left(\frac{K_t^{b,r}}{B_t^r}\right)^2 - \left(1 - \alpha^b\right) \kappa_{KB} \left(\frac{K_t^{b,g}}{B_t^g} - \nu^{b,g}\right) \left(\frac{K_t^{b,g}}{B_t^g}\right)^2 \right) \tag{79}$$

[56] Adjustment costs take the following form: $\frac{\kappa_l}{2} \left(\frac{b^{l,j}(r_t^l)}{b_{ss}^{l,j}} - 1\right)^2$

[57] The market share for loans, $\alpha^B = \left(\frac{\nu^{b,g}}{\nu^{b,r}} \frac{\alpha^b}{(1-\alpha^b)}\right) / \left(1 + \frac{\nu^{b,g}}{\nu^{b,r}} \frac{\alpha^b}{(1-\alpha^b)}\right)$ and deposits, $\alpha^d = \left(\frac{1-\nu^{b,g}}{1-\nu^{b,r}} \frac{\alpha^B}{(1-\alpha^B)}\right) / \left(1 + \frac{1-\nu^{b,g}}{1-\nu^{b,r}} \frac{\alpha^B}{(1-\alpha^B)}\right)$.

Based on empirical findings on bank capital ratios, I established that $\nu^{b,r} > \nu^{b,g}$. Consider the partial effect of changes in the extensive margin, α^b, on the aggregate loan rate:

$$\frac{\partial \bar{r}_t^l}{\partial \alpha^b} = \kappa_{KB} \frac{\epsilon^l}{(\epsilon^l - 1)} \left\{ - \left(\frac{K_t^{b,r}}{B_t^r} - \nu^{b,r} \right) \left(\frac{K_t^{b,r}}{B_t^r} \right)^2 + \left(\frac{K_t^{b,g}}{B_t^g} - \nu^{b,g} \right) \left(\frac{K_t^{b,g}}{B_t^g} \right)^2 \right\} < 0. \tag{80}$$

Equation (80) reveals that the sign depends on the difference between $\left(\frac{K_t^{b,r}}{B_t^r} - \nu^{b,r}\right)$ and $\left(\frac{K_t^{b,g}}{B_t^g} - \nu^{b,g}\right)$. Recall from Section 6.2.2 that the arguments in parentheses are positive in response to a monetary tightening, as aggregate lending declines more than bank capital, and the gap widens relatively more the smaller ν^b. The term in parentheses on the right outweighs the left, and the difference in curly brackets positive. Therefore, an increased share of regional banks lowers the loan rate (and pass-through). The magnitude depends on the relative difference between $\nu^{b,r}$ and $\nu^{b,g}$, which enhances composition effects. □

Proof: Proposition 4. Consider the partial effect of changes in the intensive margin, $\nu^{b,g}$, on the aggregate loan rate:

$$\frac{\partial \bar{r}_t^l}{\partial \nu^{b,g}} = -\frac{\epsilon^l}{(\epsilon^l - 1)} \left(1 - \alpha^b\right) \kappa_{KB} \left[\left(\frac{\mathrm{d}\left(\frac{K^{b,g}}{B^g}\right)}{\mathrm{d}\nu^{b,g}} - 1 \right) \left(\frac{K_t^{b,g}}{B_t^g} \right)^2 + 2 \frac{\mathrm{d}\left(\frac{K^{b,g}}{B^g}\right)}{\mathrm{d}\nu^{b,g}} \left(\frac{K_t^{b,g}}{B_t^g} - \nu^{b,g} \right) \right] < 0 \tag{81}$$

Since the cross-partial $\frac{\mathrm{d}\left(\frac{K^{b,g}}{B^g}\right)}{\mathrm{d}\nu^{b,g}}$ is positive, all terms are positive, and an increase in the giant bank's capital ratio, $\nu^{b,g}$, leads to a lower rate. Same principle as before, the effect depends on the size of the market share, α^b, which lowers the effect. □

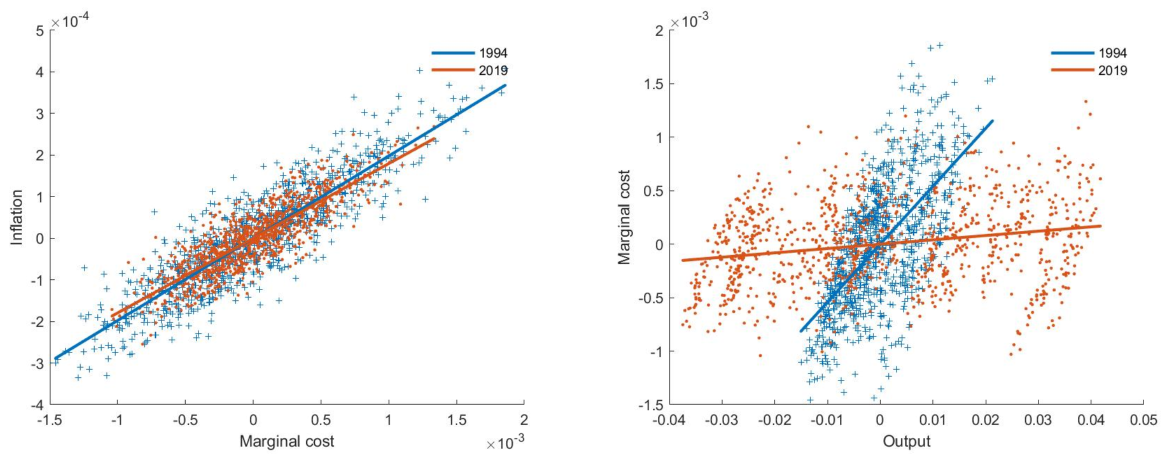
B.8 Disentangling the Flattening of the Phillips Curve

This section aims to shed light on the flattening of the Phillips curve. I decompose the total shift in the Phillips curve into three components: inflation, output, and marginal costs. Figure B.4 reveals that the break between the relationship of output and inflation is mostly due a break between marginal costs and output and not between inflation and marginal costs.

Figure B.4: Relation between inflation, marginal costs and output

(a) Inflation and marginal costs

(b) Output and marginal costs



Notes: Simulated data for output, marginal costs, and inflation based on banking sector calibration to 1994 and 2019. Data expressed in terms of deviations from steady state (unconditional mean).

B.9 Extension Borrowing Constraints on the Household and Firm Side

This section examines the role of bank concentration for monetary policy pass-through in an environment where households and firms face financial frictions. Financial frictions are an important factor – with about 31% of households in the US being borrowing-constrained (Grant, 2007). An LTV-ratio restricts most mortgage and investment loans. In the case of mortgages, the maximum loan volume corresponds to a fraction of the housing value. In this extension, the impatient household faces a borrowing constraint à la Iacoviello (2005) and the entrepreneur a borrowing constraint connected to the physical capital, shown in equations (82) and (83). The impatient household's borrowing amount, $\left(1+r_t^{bH}\right) b_t^I$, is limited by a maximum LTV-ratio, $\epsilon^{m,I}$, tied to the housing stock, h_t^I, times the expected future house price, $\mathbb{E}_t q_{t+1}^h$, and expected future inflation, $\mathbb{E}_t \pi_{t+1}$. Similarly, the entrepreneur's borrowing amount, $\left(1+r_t^{bE}\right) b_t^E$, is restricted by a maximum LTV-ratio, $\epsilon^{m,E}$, times the depreciated capital stock, $(1-\delta)\, k_t^E$, the expected price of capital, $\mathbb{E}_t q_{t+1}^k$, and the expected future inflation rate, $\mathbb{E}_t \pi_{t+1}$.

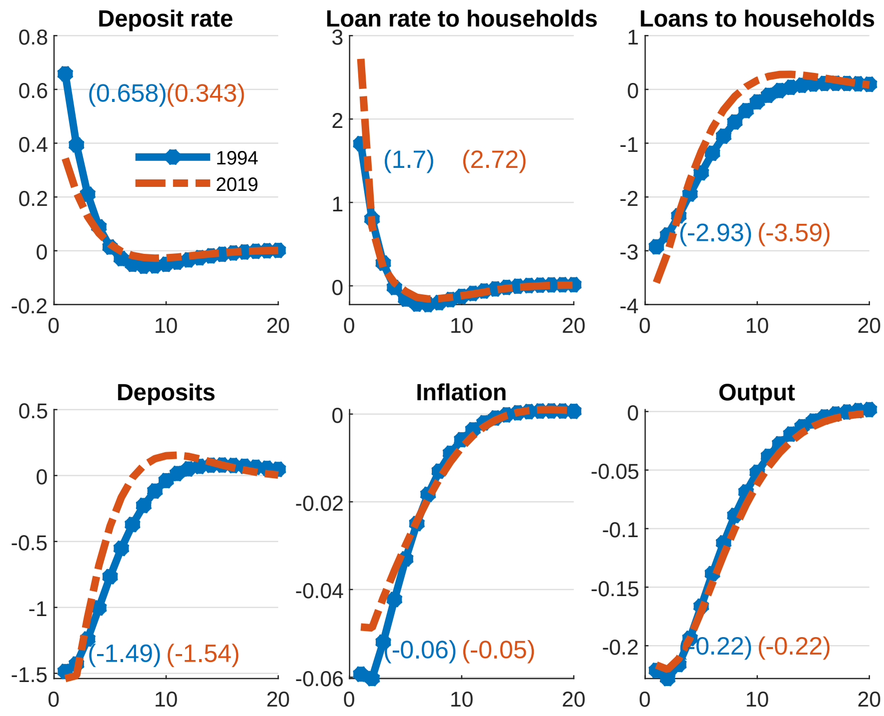
$$\left(1+r_t^{bH}\right) b_t^I \leq \varepsilon^{m,I} \mathbb{E}_t \left[q_{t+1}^h h_t^I \pi_{t+1}\right] \tag{82}$$

$$\left(1+r_t^{bE}\right) b_t^E \leq \varepsilon^{m,E} \mathbb{E}_t \left[(1-\delta) q_{t+1}^k k_t^E \pi_{t+1}\right] \tag{83}$$

This modification leads to a financial accelerator effect: a monetary tightening leads to a more severe economic downturn (i.a., lower inflation, output, and asset prices) as collateral constraints tighten and loan demand declines independently of higher interest costs. Consequently, this decreases the agent's interest-rate sensitivity, i.e., making the agents less sensitive to changes in the loan rate.

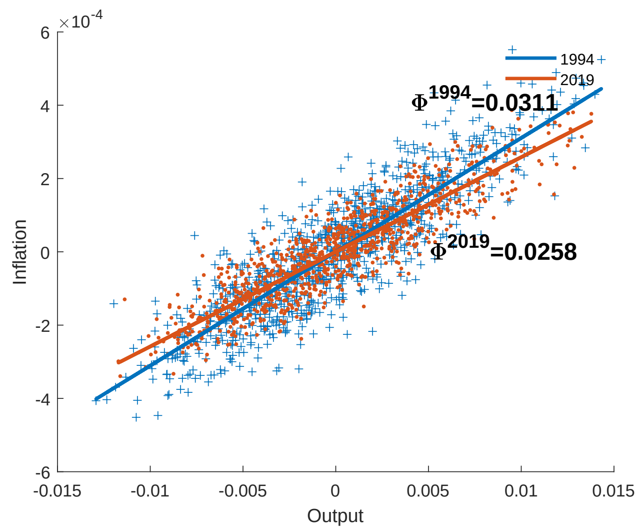
Figure B.5 compares the impulse response functions of deposit and loan rate, deposits, household loans, output, and inflation to a monetary shock in a banking environment of 1994 and 2019. The impulse response functions of the loan and deposit rates are qualitatively similar to Figure 17. Therefore, adding borrowing constraints does not significantly alter the pass-through to interest rates. However, there are different effects on the credit cycle. Loans and deposits are more responsive in 2019 versus 1994, though the difference is smaller than seen in the unconstrained model. Further, the effect on inflation is more muted in 2019, similar to the unconstrained model, but the difference is smaller. The response of output is unaltered from bank concentration in this environment. However, the reduced effect on inflation still leads to a flatter observed Phillips Curve over time shown in Figure B.6.

Figure B.5: Impulse responses to a monetary tightening varying α^b, α^m, ϵ, and ν^b



Notes: Impulse responses to a positive monetary shock in 1994 (2019) in solid blue with asterisks (red-dashed). The difference between 1994 and 2019 are shifts in α^b, α^m, ϵ, and ν^b. The impact effect is displayed in parentheses.

Figure B.6: Phillips curves: relation between inflation and output



Notes: Simulated data for output and inflation based on banking sector calibration to 1994 and 2019. Data expressed in terms of deviations from steady state (unconditional mean).