Villacorta, Alonso

Working Paper

Business cycles and the balance sheets of the financial and non-financial sectors

ESRB Working Paper Series, No. 68

Provided in Cooperation with:
European Systemic Risk Board (ESRB), European System of Financial Supervision

Suggested Citation: Villacorta, Alonso (2018) : Business cycles and the balance sheets of the financial and non-financial sectors, ESRB Working Paper Series, No. 68, ISBN 978-92-9472-020-7, European Systemic Risk Board (ESRB), European System of Financial Supervision, Frankfurt a. M., https://doi.org/10.2849/111676

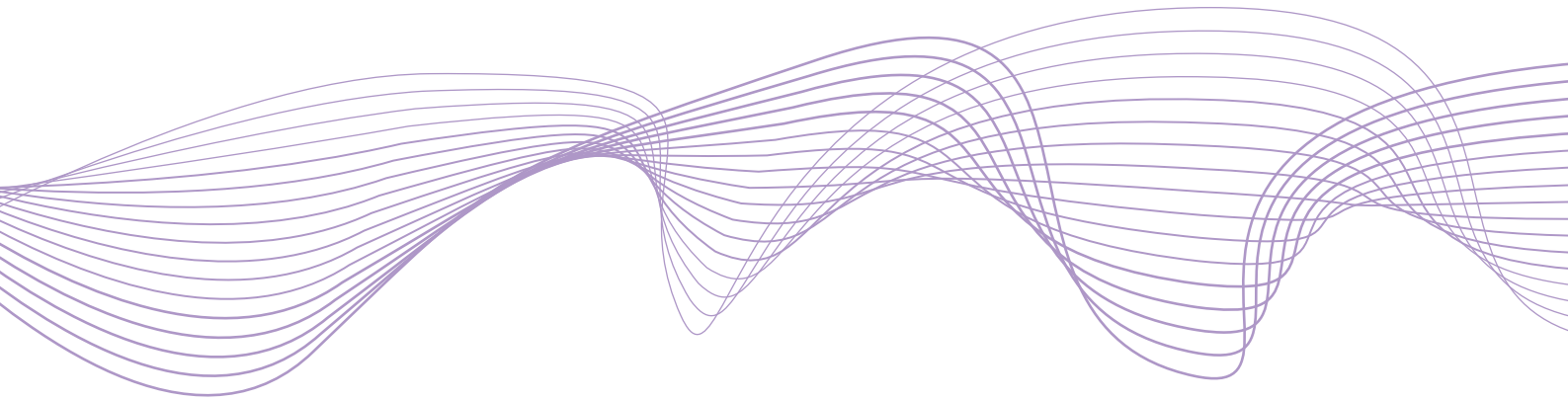
This Version is available at:
https://hdl.handle.net/10419/193575

Working Paper Series

No 68 / February 2018

Business cycles and the balance sheets of the financial and non-financial sectors

by
Alonso Villacorta



ESRB
European Systemic Risk Board
European System of Financial Supervision

Abstract

I propose and estimate a dynamic model of financial intermediation to study the different roles of the condition of banks' and firms' balance sheets in real activity. The net worth of firms determines their borrowing capacity both from households and banks. Banks provide risky loans to multiple firms and use their diversified portfolio as collateral to borrow from households. This intermediation process allows additional funds to flow from households to firms. Banks require net worth for intermediation as they are exposed to aggregate risk. The net worth of banks and firms are both state variables. In normal recessions, firm and bank net worth play the same role, so their sum determines the allocation of capital. During financial crises, shocks to bank net worth have an additional effect beyond that in standard financial frictions' models. This mechanism works through intermediation and affects activity, even if shocks redistribute net worth from banks to firms. I estimate my model and find that the new mechanism accounts for 40% of the fall in output and 80% of the fall in bank net worth during the Great Recession. Finally, the model is consistent with the different dynamics of the share of bank loans in total firm debt and credit spreads during the recessions of 1990, 2001, and 2008.

Keywords: Financial Frictions, Financial Markets and the Macroeconomy, Financial Crises, Balance Sheet Channel.
JEL codes: E44, E32, G01.

1 Introduction

During the 2008 financial crisis, banks suffered large losses to their asset values which led to a banking crisis and massive bailouts, followed by a long-lasting recession. This paper analyzes the importance of the health of banks' balance sheets for the severity of economic recessions. I ask, are banks "special" in the sense that shocks that affect banks' balance sheets matter more for real economic activity than shocks that affect other firms? For the policy debate, it is crucial to understand the mechanisms by which shocks that affect banks or firms propagate to real activity. For instance, is recapitalizing banks more effective than recapitalizing other borrowers during recessions?

There is empirical evidence that both shocks that affect firm net worth (firm credit channel[1]) and shocks to bank net worth (bank lending channel[2]) propagate through balance sheet restrictions and affect activity. However, few general equilibrium frameworks consider both channels simultaneously. This paper provides a model in which both channels are present and can quantitatively capture the differential dynamics of macroeconomic and financial variables in different recessions. Having a framework that quantitatively considers both endogenous channels is important to speak to whether banks are "special" and to guide the policy response in the most effective way (depending on the relative importance of each channel).

To motivate the analysis, Figure 1 shows two recessions that differentially affected the financial and the non-financial sectors. Figure 1 displays the evolution of aggregate net worth, as measured by the market value of equity, in the financial and non-financial sectors as a ratio of GDP during the recessions of 2001 and 2008. The whole corporate sector experienced large wealth losses in both recessions (panel A); however, the consequences on real activity were different the recession of 2001 was shorter and milder. Panel B shows that while the 2001 recession severely affected the non-financial sector, it mildly affected financial sector balance sheets, which instead severely deteriorated during the Great Recession.[3]

Moreover, the behavior of the bank lending market was different in the two recessions. Figure 2 displays the share of bank loans in total firm debt (panel A) and the cost of bank loans (panel B). Both series show different dynamics during these two recessions. The share of bank loans experienced a substantial drop during the 2008 recession while a small change in 2001. The cost of bank loans shows a large increase during the Great Recession while a

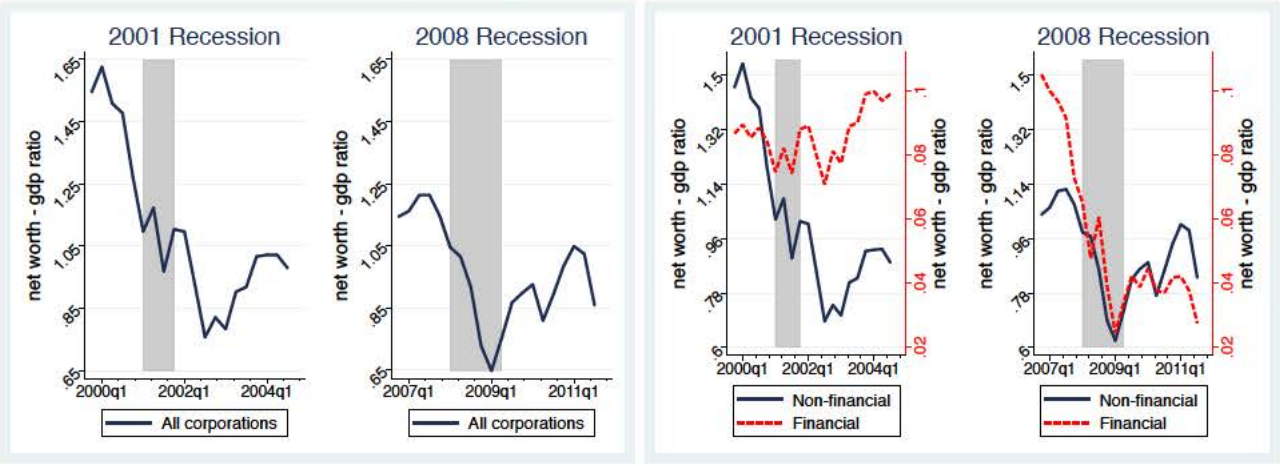
[1]See Hubbard [1998] and Stein [2003] for surveys; and Giroud and Mueller [2015] for recent evidence.

[2]See Bernanke and Blinder [1992], Kashyap and Stein [1994], Bernanke and Gertler [1995], Kashyap and Stein [2000], Ivashina and Scharfstein [2010] and Chodorow-Reich [2013].

[3]In a 2012 speech, Fed Chairman Bernanke stressed: "any theory of the crisis that ties its magnitude to the size of the housing bust must also explain why the fall of dot-com stock prices, which destroyed as much or more paper wealth - more than $8 trillion - resulted in a relatively short and mild recession". See Ben S. Bernanke (2012), "Some Reflections on the Crisis and the Policy Response", at the Russell Sage Foundation and the Century Foundation Conference.

Figure 1: Net Worth in the 2001 and the 2008 Recessions

(a) Net Worth (All Corporations) (b) Net Worth (Financial and Non-financial)



Note. Panel (b): Market value of equity of non-financial corporate businesses as a ratio of GDP (solid blue line) and market value of equity of U.S. commercial banks as a ratio of GDP (connected red dots). Panel(a): All Corporations is the sum of the two sectors. Source: US Flow of Funds. Shaded areas indicate NBER-dated recessions.

small change in 2001. This differential response of the market of bank lending suggests that the mechanisms in place during these two recessions were different.

In this paper, I develop and estimate a framework to investigate the mechanisms that emerge in different recessions: when banks' balance sheets are more affected (e.g. 2008) or when firms' are more affected (e.g. 2001). I propose a dynamic model of financial intermediation with three agents: borrowers (firms) who are experts at production, lenders (households), and banks who use their monitoring skills to make risky loans to firms. Contrary to most previous work, financial frictions in all relations imply that both bank net worth and firm net worth jointly affect equilibrium dynamics. The main result is that bank net worth is "special" if and only if the share of bank net worth is sufficiently low I refer to these times as financial recessions. Thus, I uncover a new amplification mechanism active only in financial recessions by which shocks to bank net worth have an additional effect on activity beyond that generated by the standard financial accelerator, so they matter more than shocks to firm net worth. This non-linearity in the model leads to different kinds of recessions that are consistent with the findings from the empirical literature on financial crisis.[4] Moreover, banks are special and policy should treat them differently than other borrowers during financial recessions but not in regular recessions.

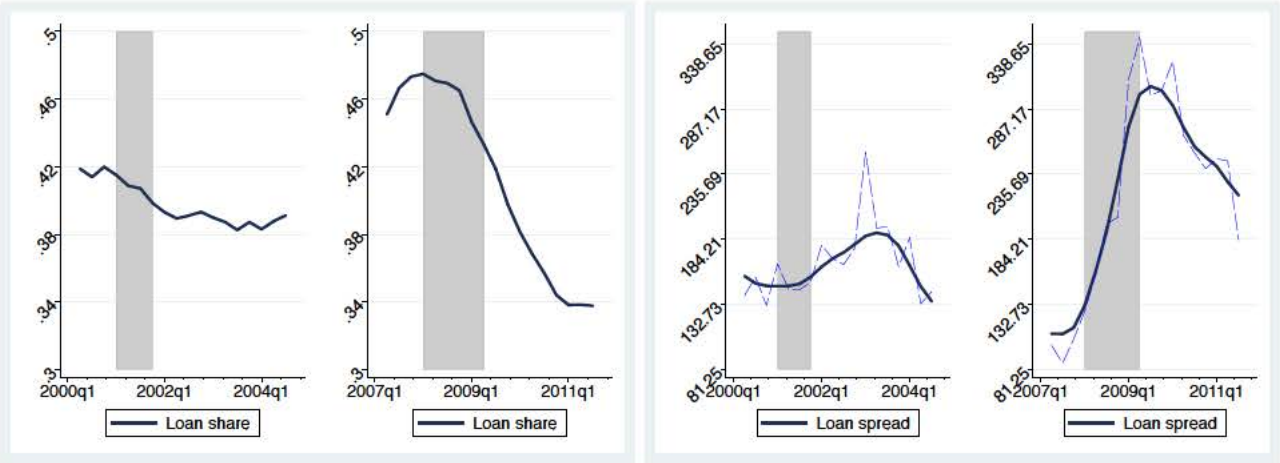
The second contribution is quantitative; I show that the model can match key macroeconomic and financial data. For instance, the early 90s and 2008 recessions which were accompanied by large drops in bank net worth relative to the 2001 recession were associated with

[4]See Related Literature (sub-section 1.1).

Figure 2: Bank Lending Market

(a) Share of Bank Loans in Total Firm Debt

(b) Cost of Bank Loans



Note. Panel (a): Share of loans in all credit market instruments (outstanding), for non-financial corporations. Source: US Flow of Funds. Panel (b): Weighted average of the cost of new loan issuances (in bps) extended for general corporate purposes and liquidity management. Cost is defined as all-in-drawn spread, which is total (interest plus recurring fees) spread paid over 6 month LIBOR. Source: DealScan database of loan originations. Panel reports the raw series (dashed line) and its smoothed version (solid line).

steeper and more persistent declines in output and in the share of bank loans in firm debt. The estimation finds that the new non-linear mechanism was active in those "financial" recessions but not in the recession of 2001. Therefore, the model can explain these key stylized facts. Moreover, the model can match a number of overidentifying restrictions. For instance, the model succeeds in explaining the untargeted behavior of credit spreads and the share of bank loans during these three recessions.

To illustrate how the mechanism works, it is helpful to consider the model in more detail. In the model, both firms and banks face moral hazard problems that lead to endogenous borrowing constraints. Firms have access to two sources of financing: directly from households or intermediated loans from banks. But, due to moral hazard problems, firm net worth constrains the amount of borrowing from both types of financing (debt must be secured by collateral). Households only accept riskless debt (secured by safe collateral). Banks are special in their monitoring skills, which allows them to make firms repay in good states, and thus extend risky loans (accept risky collateral). So, intermediated finance, while subject to credit risk, allows firms to raise additional funds. By lending to multiple firms, banks pool idiosyncratic risks and use their diversified portfolio as collateral to borrow additional funds from households. This intermediation process diversification and creation of collateral increases the available resources for investment. However, banks' portfolios are exposed to aggregate shocks. So, banks require net worth for intermediation to secure their debt from households. In this model, both firm

net worth and bank net worth are state variables that matter for economic activity and can potentially have a differential effect.

The economy is characterized by different regimes that are endogenously defined by the relative size of bank and firm net worth. The behavior of the market of bank lending (equilibrium loan spread and volume), the response of aggregates to shocks and the relative importance of bank and firm net worth for activity is different in each regime. In particular, when the share of bank net worth is critically low (i.e. the ratio of bank to firm net worth is below a threshold), the intermediation process becomes disrupted. Therefore, the economy can enter into different kinds of recessions depending on whether bank net worth is critically low or not: (i) non-financial recessions and (ii) financial recessions.

During non-financial recessions, bank and firm net worth play the same role in affecting activity. In this regime, a drop in bank net worth tightens banks' constraints and hurts banks' ability to make loans. The reduction in the supply of loans increases the loan rate (i.e. higher spreads), which in turn helps banks. The higher cost of loans affects firm financing constraints and hurts activity. Even when firms are willing to borrow at the higher rates, the higher cost of loans reduces the maximum funds firms can get given their net worth (tighten firm constraints both with households and banks). However, the effect is similar to just a drop in firm net worth. Through the loan rate, banks and firms share the costs generated by shocks to their net worth. As a result, the effect on activity from a drop in bank net worth is equivalent to the effect from a drop in firm net worth. In this regime, what is relevant for aggregate activity is their consolidated net worth, so recapitalizing banks has similar effects than recapitalizing firms.[5]

In contrast, financial recessions are times in which bank net worth is critically low. In this regime, the cost of loans is remarkable high relative to firms' expected profitability, for instance, the loan rate is at the maximum firms are willing to accept given the return on their assets. During these times, a drop in bank net worth cannot generate a further increase in the loan rate, as firms prefer to forego expensive risky loans than to face an increase in the rates. Even when firms are constrained and borrow up to the maximum from households, they choose to borrow less risky loans than the maximum allowed by banks (firm constraint with households is binding, while constraint with banks is not). Banks could use those loans to create collateral and get additional funds from households. Thus, this reduction in lending implies a loss of intermediation drop in collateral created by banks that further constrains bank supply of loans and amplifies the initial shock. This feedback effect through the intermediation process is unique of the financial recession regime and is the source of the new amplification mechanism in the model. As a result, a drop in bank net worth has more severe consequences on real

[5]Note that this aggregation result is different from the existing literature that assumes frictionless relations between firms and banks. In this regime, financial frictions are always present in the loan market.

activity than a drop in firm net worth. This new mechanism implies that financial recessions are longer and more severe than other recessions and favors policies directed to recapitalize banks rather than firms during these times.

Different shocks can move the economy into the different regimes. In reality, we think that the assets of financial and non-financial firms are subject to different shocks, some shocks affect relatively more non-financial firms (e.g. the dot-com bust in 2001) others affect more financial intermediaries (e.g. the housing crisis in 2008). The model explains how different shocks amplify, propagate and can differentially affect the real economy, through the endogenous mechanisms, depending on which sector is affected. The estimation of the model allows for two different shocks that can move the economy into the different regimes. First, a productivity shock affects the return on assets of firms, which in turn affects bank loan repayments; thus a productivity shock affects both firm and bank net worth. Second, a bank-specific shock that affects only bank net worth; this shock tries to capture other bank business such as residential mortgage lending, which is absent in the model. While the bank shock does not directly affect the assets of the firm, it can indirectly propagate by its effects on the market of bank lending.

The second contribution of the paper is to quantify the importance of the new non-linear mechanism. I estimate the model using U.S. data on output (real GDP) and the aggregate market value of bank net worth from 1980 to 2015. The estimation is based on a maximum likelihood approach. The model is dynamic and features non-linearities related to the different regimes. I use a particle filter, as suggested by Fernández-Villaverde and Rubio-Ramírez [2007], to deal with the non-linear system.

First, given the estimated parameters and latent variables, I identify when did the U.S. enter into financial and non-financial recessions from the lens of the model. The estimation identifies the recession of the early 90s and 2008 as financial, as both are associated with large drops in the relative share of bank net worth; while the recession of 2001 is not.

Then, I quantify the importance of the non-linear intermediation mechanism which is only activated during financial recessions by comparing the dynamics of the model relative to an alternative where this intermediation mechanism is shut down only the standard financial accelerator remains. In this alternative exercise, the flow of credit through the financial sector is not disrupted, for example as occurred in the 2001 recession. I find that the intermediation mechanism explains on average 40% of the fall in output and 80% of the fall in bank net worth during the Great Recession. The mechanism also induces persistence and implies longer recessions associated with financial crises. Thus, the endogenous dynamics in the financial recession regime can account for output and bank net worth remaining below their trend for several years after the Great Recession. As a result, the average one-year forecast errors increase, by 20% in the case of output and 40% in the case of net worth, when the new mechanism is not

considered.

Finally, the model can explain the differential dynamics observed in the share of bank loans in total firm financing and the cost of credit during the recessions of 1990, 2001, and 2008. These series are not used in the estimation; thus they serve as external validation of the model. First, the estimated model generates much larger drops in the share of bank loans in the recessions of 1990 and 2008, where the intermediation mechanism is active, relative to the recession of 2001. These dynamics are consistent with the observed share of bank loans in firm borrowing.

Second, the model can also explain the dynamics of the cost of credit in these three recessions. Credit spreads experienced a pronounced increase in 2008, a considerable but lower increase in 2001, and a small change in 1990. The estimated model can capture these three facts. Interestingly, the estimated model can produce such differential response in both identified financial recessions: the large increase in the 2008 recession and the small change in the early 90s. These two financial recessions are different because of two reasons. First, the severity of the recession is different, the drop in bank net worth is much larger in 2008. Second, in the model, what matters to activate the new mechanism is the loan rate relative to the return on assets of firms. The estimation finds that the early 90s recession was associated with a relatively larger productivity shock that reduced firm expected returns, whereas the Great Recession was mainly triggered by a specific shock to banks. Thus, in the early 90s, a small increase in the loan rate was sufficient to make firms contract their demand of bank loans below their limits, which triggers the mechanism that can simultaneously explain the drop in bank loans and the persistence in output and net worth.

1.1 Related Literature

This paper builds on the banking literature that stresses the monitoring role of banks, as for example Diamond [1984], Williamson [1986] or Krasa and Villamil [1992]. In these papers, financial intermediation arises endogenously as the dominant vehicle for borrowing and lending due to duplicative monitoring costs. These papers analyze the optimal contracting environment in which banks appear as delegated monitors, but there is no role for bank net worth in intermediation. Papers that have highlighted the importance of bank net worth for their intermediation activities include Holmstrom and Tirole [1997] and Diamond and Rajan [2000]. Banks are modeled as relationship lenders, with a special ability to extract repayment from borrowers. In particular, in Holmstrom and Tirole [1997], both the net worth of firms and banks matter for investment and spreads. All of these models are static, the net worth of agents is exogenous, and they do not focus on its dynamic relationship with the macroeconomy.

Dynamic models in which the net worth of borrowers plays a role are pioneered by Bernanke and Gertler [1989] and Kiyotaki and Moore [1997]. Bernanke et al. [1999] embed those fea-

tures in a macroeconomic framework suitable for quantitative analysis, in which constrained borrowers are represented by entrepreneurs in the non-financial sector. These models feature the financial accelerator as propagation mechanism, where the net worth of borrowers the productive agents matters for dynamics. Exposure to aggregate risk by those levered agents can lead to balance sheet recessions: shocks that affect borrowers' net worth are amplified as these agents become less able to hold assets and invest, depressing activity. In addition, transitory shocks persist as net worth takes time to recover. Gertler and Kiyotaki [2010], Gertler and Karadi [2011], He and Krishnamurthy [2013], Brunnermeier and Sannikov [2014] and others consider an intermediary sector.[6] These papers assume that there are no frictions between banks and firms, so that only the sum of bank and firm net worth the consolidated borrowing sector matters for the dynamics. These models provide a good understanding of why balance sheets play a role in an economy where financial frictions limit the availability of funds to borrowers. However, they don't provide a good explanation of what is particularly special about banks' balance sheets and financial recessions. Thus, they cannot address whether shocks to banks matter more for activity than shocks to other borrowers and what a redistribution from other borrowers to banks would entail.

Iacoviello [2015] includes financial frictions between banks and firms. In his model, entrepreneurs can only borrow from banks without banks, entrepreneurs cannot get any outside funding. The estimation of the model attributes 2/3 of the decline in GDP during the Great Recession to financial shocks. In the estimation, constraints always bind. Instead, my paper documents that non-linearities are important in an estimated model in which constraints only bind occasionally. These non-linearities are particularly relevant to explain the empirical differences between financial and non-financial recessions and help in the forecasting analysis. Moreover, to be able to speak about the composition of firm debt, I allow firms to borrow from banks as well as other lenders. I find that frictions in the bank lending market are important even when firms are allowed to borrow directly from households.

Two recent papers also study the importance of bank and firm net worth with non-linear dynamics and occasionally binding constraints. Elenev, Landvoigt, and Van Nieuwerburgh [2017] provide evidence from a calibrated model in which firms can only borrow from banks (as in Iacoviello [2015]). The paper uses the model to study macroprudential policy. It finds that restrictions on bank leverage promote stability but shrink the size of the economy and generate welfare losses. Rampini and Viswanathan [2017] develop a theoretical framework with endogenous constraints in which both bank and firm net worth jointly determine economic dynamics. In their model, banks are better able to enforce collateralized claims than households, so they can lend more to firms. But the additional amount banks can lend has to be financed

[6] See Brunnermeier, Eisenbach, and Sannikov [2012] for an overview of the literature of macroeconomic models with financial frictions.

out of their own net worth. A key difference in my model is that banks, by pooling risk and creating collateral, channel additional funds from households to firms. This intermediation channel makes bank net worth special and is the source of the new feedback mechanism in the model. Thus, I identify a new non-linear mechanism related to intermediation and I show that it is quantitatively important by estimating my model. Elenev, Landvoigt, and Van Nieuwerburgh [2017] and Rampini and Viswanathan [2017] generate slower recoveries from recessions that are accompanied by a credit crunch. These papers do not analyze the different effects of recapitalizing banks or firms.

In other papers that consider occasionally binding constraints (e.g. Elenev, Landvoigt, and Van Nieuwerburgh [2017], He and Krishnamurthy [2013] and Guerrieri and Iacoviello [2017]), firm constraints bind during recessions. Instead, in my framework, firms face two types of constraints: on their borrowing from banks and households. Importantly, during financial recessions, the firms' constraints with banks are slack, while only their constraints with households bind. The fact that firms don't borrow up to the maximum affects banks' ability to create collateral and induces the new amplification mechanism in the model.

This paper is also related to the empirical literature on financial crises, for instance, Bordo, Eichengreen, Klingebiel, and Martinez-Peria [2001], Cerra and Saxena [2008], Reinhart and Rogoff [2009], Claessens, Kose, and Terrones [2010], Calvo, Coricelli, and Ottonello [2012], Jordà, Schularick, and Taylor [2013] or Boissay, Collard, and Smets [2013]. These empirical studies are based on different methods to identify financial crises episodes and study their economic consequences. This literature generally concludes that recessions accompanied by financial crises are more severe and persistent than other recessions. More recently, Romer and Romer [2015] and Krishnamurthy and Muir [2016] suggest that it is important to distinguish financial recessions by their severity, so they use a continuous measure of financial distress; the former derives a measure using a narrative approach based on a reading of OECD accounts and the latter uses credit spreads and pre-crisis credit growth. These measures of financial distress intend to capture the rise in the cost of credit intermediation. Both papers find that recessions associated with more severe financial distress are associated with slower recoveries. My model generates both non-financial and financial recessions. The endogenous mechanism in the model can explain why are financial recessions more severe and persistent, and associates the severity of the downturn with the cost of credit intermediation. Thus, this paper provides a theory that supports the empirical literature. Moreover, I provide a structural framework to identify financial crises episodes. For instance, the estimation identifies the recessions of 1990 and 2008 as financial recessions and identifies the 2008 recession as being more severe.

This paper is also related to the literature that studies the relation between credit spreads and economic activity. Gilchrist and Zakrajšek [2011] and López-Salido, Stein, and Zakrajšek

[2016] show that movements in credit spreads have substantial explanatory power for future economic activity. Moreover, these papers suggest that the relevant variation in credit spreads is not related to firms' default probability but instead to risk premia and that it reflects a contraction of credit supply instead of credit demand. Adrian, Moench, and Shin [2010a], Muir [2014] or Adrian, Etula, and Muir [2014] document that fluctuations in risk premia are related to financial intermediaries' balance sheets. These papers suggest that the health of intermediaries' balance sheets determines their effective risk-bearing capacity, which in turn influences the supply of credit, risk premia, and real activity. My paper provides a theory in line with these findings. Moreover, the model implies that what matters for activity is not the variation in the level of credit spreads but its value relative to the return on firm assets. For instance, this is important to explain the dynamics in the early 90s recession, an event that is missed by standard intermediary-based models as shown by He and Krishnamurthy [2014].

Layout. The rest of the paper is organized as follows. Section 2 describes the main equations that drive the dynamics of aggregates used in the estimation. Section 3 describes the model. Section 4 solves the equilibrium dynamics in the regime of normal booms and recessions. Section 5 solves for the dynamics during the financial crisis regime. Section 6 discusses the estimation of the model. Section 7 presents the results of the quantitative exercise. Section 8 concludes.

2 The role of bank net worth: a stylized model

In this section, I present the main equations that drive the dynamics of aggregate variables in my model. The purpose of this section is to provide the intuition behind the main equations used in the estimation without deriving them from first principles.

Financial frictions imply a relation between investment and the net worth of borrowers. For example, in models with collateral constraints, net worth determines the collateral value that firms can provide to their creditors. Thus, models of financial frictions feature dynamics that are represented by

$$K_t \leq \lambda_t N_t, \tag{1}$$

where K_t is aggregate capital, N_t is aggregate net worth of borrowers and λ_t is a multiplier associated with financial constraints; when the constraint is binding λ_t represents leverage.[7]

Borrowers invest $1/\lambda_t$ per unit of capital from their own net worth and the rest, $B_t = K_t - N_t$, is the value of debt borrowed from creditors. Denote R as the promised return to creditors and

[7]For example, in Kiyotaki and Moore [1997], λ_t is associated with the future price of capital q_t because this price determines the value of collateral: $\lambda_t = (q_t - q_{t+1}/R)^{-1}$. Instead, without financial constraints, we have $\lambda_t \to \infty$.

R^K_{t+1} as the return on capital, then

$$N_{t+1} = R^K_{t+1} K_t - RB_t. \tag{2}$$

Equations (1) and (2) show the dynamic complementarities between capital and net worth, which are the main forces behind the so called "financial accelerator". Shocks that affect net worth N_t affect investment and output, which in turn affect future values of net worth, capital, and output. Through this mechanism, even temporary shocks can generate cycles.

Both banks and firms are borrowers that, in the end, invest and share the risks and profits of real projects, while facing frictions from other creditors (households). Thus, the net worth that matters is the sum of net worth of banks and firms and this is what N_t represents.[8]

It is common to associate the degree of financial frictions, captured by λ_t, with the functioning of the financial sector. Banks act as intermediaries which alleviate frictions and help funds to flow from creditors to borrowers. During normal times, funds flow from creditors to firms through banks, implying high levels of borrowing and investment per unit of net worth (λ_t high). During financial recessions, banks intermediation process is disrupted and credit per unit of net worth falls, implying a reduction in borrowing and investment (λ_t low).

Comparison to previous literature. Christiano, Motto, and Rostagno [2014] and Liu, Wang, and Zha [2013] perform a business cycle accounting exercise in models with financial frictions and financial shocks. They find that financial shocks are important in explaining cyclical variations of investment and output, and are particularly important in explaining the observed fluctuations during the Great Recession. Their models consider relations that can be represented by (1) and (2), in which financial shocks are associated to exogenous variations of λ_t. Christiano, Motto, and Rostagno [2014] interpret financial shocks as uncertainty shocks which imply changes in the contracts between firms and creditors. In particular, when uncertainty is high, risk-sharing between firms and creditors falls, which leads to a decrease in λ_t. Liu, Wang, and Zha [2013] interpret financial shocks as "collateral shocks" that are direct shocks to borrowing constraints (exogenous shocks to λ_t). My model explains financial crises by generating endogenous changes in λ_t due to adjustments of banks' balance sheets.

In my model, due to their monitoring skills banks are able to share risks with firms that other creditors are not willing/able to share. By lending to multiple firms, banks pool idiosyncratic risks and use their diversified portfolio as collateral to borrow from households. Through this intermediation process banks receive additional funding that in turn is lent to firms. Thus, intermediation determines the supply of credit and investment by firms. As a result, banks'

[8] For example, Gertler and Kiyotaki [2010] calibrate their model to match an average leverage ratio across financial and non-financial sectors.

intermediation process determines the multiplier λ_t.

Banks are simultaneously borrowers and lenders, and have frictions on both sides of their balance sheets. Banks diversify idiosyncratic risk but hold aggregate risk in their balance sheets and so they require net worth for intermediation, as they have to repay their debt in case of a negative aggregate shock. Thus, the net worth of banks determines the tightness of their balance sheets constraints and their debt capacity, which in turn determines the supply of funds intermediated to firms.[9] As a result, my model implies a relation between the multiplier λ_t and bank net worth. Denoting with β_t the share of bank net worth relative to total net worth of borrowers (banks and firms), i.e. $\beta_t = \frac{N_t^B}{N_t}$, I capture these relations with the following model for λ_t

$$\lambda_t = \begin{cases} \bar{\lambda} & \text{if } \beta_t \geq \bar{\beta} \\ \lambda(\beta_t) & \text{if } \beta_t < \bar{\beta} \text{ (financial recessions)} \end{cases} \tag{3}$$

where $\lambda(\bar{\beta}) \leq \bar{\lambda}$ and $\lambda(\beta_t)$ depends positively on β_t. This specification associates financial recessions with bank net worth losses (weak balance sheets). Moreover, the specification associates the severity of financial recessions with the magnitude of net worth losses. Thus, the persistence and severity of the recession depends on the evolution of β_t. The evolution of bank net worth depends on the returns of their investments, which consist of loans to firms. Denote the representative portfolio of bank loans as L_t and its corresponding return as R_{t+1}^B (loan rate). Through loans, banks invest on firm projects and share firm risks and returns. Then, the evolution of bank net worth is represented by

$$N_{t+1}^B = R_{t+1}^B L_t - RD_t, \tag{4}$$

where D_t is the aggregate debt of banks (deposits). The evolution of β_t follows from equations (2) and (4).

The model described by equations (1)-(4) allows me to differentiate recessions where firms balance sheets are affected and so total net worth N_t drops (e.g. the 2001 recession) from recessions where, in addition, financial firms are severely affected and so the share of bank net worth, β_t, drops (e.g. the 2008 recession).

During financial recessions, banks' balance sheet constraints tighten, loan rates increase, and the supply of loans contracts. Therefore, the model is also able to produce dynamics for the volume of bank loans and for the loan rate that vary across financial and other recessions. For instance, the composition of firm debt, from banks or other creditors, changes with β: the share of bank loans in total firm debt decreases more during financial recessions relative to

[9] Adrian et al. [2010b] refer as banks "risk appetite" to this idea that relates banks' balance sheet constraints, banks' risk-bearing capacity and the supply of credit.

other recessions; while loan rates increase relatively more. Denoting the share of bank loans in total firm debt with $share_t^L$, the model implies functions

$$R_{t+1}^B = r^B(N_t, \beta_t),$$

$$share_t^L = s^B(N_t, \beta_t),$$

where $r^B(.)$ is decreasing in β_t and $s^B(.)$ is increasing in β_t. In the next section, I microfound the equations presented in this section.

3 Model

In this section I develop a simple model where financial frictions restrict risk sharing and constrain the flow of funds to the productive sector.

I consider a dynamic economy that is populated by three types of risk neutral agents: a continuum of households, entrepreneurs and banks. Time is discrete and the horizon infinite. There is one good that can be used for consumption or investment in capital.

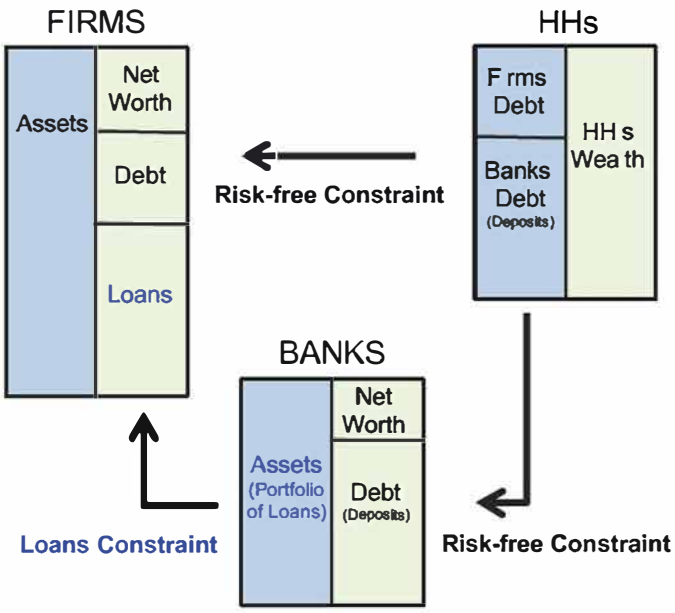
Entrepreneurs have access to productive projects but their net worth is limited. Entrepreneurs privately observe projects' returns. This informational friction constrains the flow of funds to entrepreneurs. Households do not share the risks of the projects, so entrepreneurs only raise funds from households through riskless debt. Banks, unlike households, can monitor entrepreneurs and are able/willing to share additional risks. Monitoring outcomes are also privately observed by the bank (not by households), thus households don't share risks with banks, either. In addition, monitoring is not perfect which implies a limited amount of risk-sharing between banks and firms. A description of the flow of funds restrictions is shown in Figure (3).

I now describe a simple environment where financial frictions arise as part of the contracting problem with moral hazard. The assumptions chosen are intended to simplify the exposition. The important feature of the environment is that there exist risk-sharing restrictions between borrowers and lenders that limit financing, and banks are able/willing to share and pool additional risks.

3.1 Households

There is a continuum of households of measure unity in the economy. They are risk neutral (linear utility) and discount the future by the factor $1/R$. Households are unconstrained agents assumed to have large wealth (deep pocket investors) and are willing to offer any amount of capital at an expected rate of return R. However, households are uninformed agents, they cannot monitor and enforce payments, so they don't hold risks in their contracts with firms

Figure 3: Flow of funds constraints



nor banks. As explained below, because of an informational problem, they only buy risk-free assets: bank deposits or riskless firm bonds. The riskless rate R can also be interpreted as the return in a low-productivity project to which all agents have access.

3.2 Firms (entrepreneurs)

There is a continuum of agents who run firms, which can be thought of as owners, managers or entrepreneurs. Throughout the paper, I say entrepreneurs or firms to refer to those agents. Entrepreneurs are indexed by $i \in [0,1]$. Entrepreneurs have access to the following productive technology: for each $k_{i,t-1}$ units of capital invested at time "$t-1$", entrepreneur "i" produces

$$y_{i.t} = z_{i.t} A_t k_{i.t-1}$$

where A_t represents the common technology to all entrepreneurs and $z_{i,t}$ denotes an idiosyncratic productivity shock which is privately observed and distributed $Bernoulli(p_t)$. Thus, a fraction p_t of firms fail on their projects and have zero production.

From the perspective of each entrepreneur, the production technology has constant returns, however $A_t = A(K_{t-1})$ depends on the aggregate capital K_{t-1} with decreasing returns.[10]

Capital depreciates at rate δ. In addition, a fraction κ of the depreciated capital is also affected by the shock $z_{i,t}$, e.g. firms that fail on their projects also lose part of the capital

[10]For example, decreasing returns appear in the standard framework with a Cobb-Douglas production function and a competitive labor market. See the appendix A.

invested. Therefore, the return on capital of firm "i" is

$$R^k_{i,t+1} = \begin{cases} A_{t+1} + (1-\delta) & \text{with prob. } p_{t+1} \\ (1-\kappa)(1-\delta) & \text{with prob. } 1-p_{t+1} \end{cases} \tag{5}$$

The fraction of firms that succeed in their projects ($z_{i,t+1} = 0$) is stochastic. The probability p_{t+1} is i.i.d. and follows a distribution F_p with mean $\bar{p}$ and support $(p_L, p_H]$, with $p_L > 0$. The probability p_{t+1} is the only source of aggregate shocks in the economy. The common return A_{t+1} is known at t.

Entrepreneurs start period t with net worth $n^E_{i,t}$ and can raise financing from both households and banks.

3.2.1 Firms' borrowing constraints

I restrict the analysis to short term contracts. A contract at t specifies the amount borrowed by the entrepreneur and the payments at $t+1$ as a function of the entrepreneur's reports on returns at $t+1$. No future or past reports can be used. [11]

A moral hazard problem creates financial frictions. The realization of returns on assets are privately observed by entrepreneurs, which in absence of proper incentives may misreport cash-flows in order to enjoy private benefits. Formally, entrepreneurs can announce failure ($z_{i,t+1} = 0$) and steal part of the unobserved returns. In particular, they can steal all the cash-flows and the fraction s of the capital subject to the shock, i.e. returns on stealing are $s(\kappa(1-\delta))k_{i,t}$ if $z_{i,t+1} = 1$ and 0 otherwise. In absence of monitoring, entrepreneurs can costlessly misreport and divert funds ($s = 1$). Thus, households, which cannot monitor, don't share project risks. Entrepreneurs borrow from households only by issuing riskless debt, B_t, secured by the riskless part of firm returns (value of assets under $z = 0$) which can be interpreted as the value of the collateral of the firm. Therefore the borrowing constraint for firms' riskless debt is:

$$RB_{i,t} \leq (1-\kappa)(1-\delta)k_{i,t}. \tag{6}$$

In the spirit of Holmstrom and Tirole [1997] or Diamond [1984], banks can monitor firms and reduce the benefits associated with stealing. Under monitoring, misreporting becomes costly: for each unit of cash-flow the entrepreneur misreports, he keeps $(1-\bar{\phi})$ units and $\bar{\phi}$ is lost in this stealing process, i.e. $s = 1-\bar{\phi}$. Monitoring allows banks to share project risks as long as the traded securities enforce truthful reporting. Still, the monitoring outcome is not revealed to households: entrepreneurs privately report to each of the agents they are contracting with, and so the entrepreneur-household contracts are independent of the entrepreneur relation with

[11] See the appendix B for a more detailed description of the contracts.

banks.

Any traded security between banks and an entrepreneur consists of payments in case of success $\pi^S_{i,t+1}$ or failure $\pi^F_{i,t+1}$. The entrepreneur can promise to banks different payments in case of success or failure as long as they enforce truthful reporting, which is captured by the incentive compatibility condition (IC)

$$\pi^S_{i,t+1} - \pi^F_{i,t+1} \leq \bar{\phi}\kappa(1-\delta)k_{i,t}.$$

The binomial nature of the returns implies that the contract can be implemented with riskless debt and defaultable debt. [12,13] The recovery value of the defaultable debt is indeterminate: the payment in the low state π^F_{t+1} can represent the value of the riskless debt held by the bank or the recovery value of the risky debt in case of default. Thus, without loss of generality, I implement the contract with defaultable loans with promise rate R^B_{t+1} and zero recovery value ($\pi^F_{t+1} = 0$), while payments in case of failure are all represented with riskless debt which is accounted in $B_{i,t}$. The share of firms' debt $B_{i,t}$ held by households or banks is undetermined in the model, thus I interpret that all firm riskless debt is held by households.[14]

I denote the risky return on bank loans provided to firm i as $R^B_{i,t+1}$, which equals R^B_{t+1} in case of success and zero in case of failure. And I denote $L_{i,t}$ as the total amount borrowed through bank loans, thus the total promised payment to banks in case of success is $\pi^S_{i,t+1} = R^B_{t+1}L_{i,t}$ and the (IC) becomes

$$R^B_{t+1}L_{i,t+1} \leq \bar{\phi}\kappa(1-\delta)k_{i,t}. \tag{7}$$

Entrepreneur i starts the period with net worth $n^E_{i,t}$, borrows $B_{i,t}$ through riskless debt and $L_{i,t}$ through bank loans.[15] Therefore, total capital investment by firm i at period t is

$$k_{i,t} = n^E_{i,t} + B_{i,t} + L_{i,t}. \tag{8}$$

3.2.2 The bank loan rate and the loan premium

Riskless debt and bank loans are imperfect substitutes as sources of financing, so they can carry different prices. I denote the expected return on bank loans as $\bar{R}^B_{t+1} = \mathrm{E}_t p_{t+1} R^B_{t+1} = \bar{p}R^B_{t+1}$.

[12] Any contract can be replicated by two not perfectly correlated assets.

[13] In a similar framework but in a more general setting, for any stochastic distribution of returns and any form of monitoring costs, Krasa and Villamil [1992] show that optimal contracts are characterized by two sided simple debt: simple debt between lenders and banks and simple debt between banks and firms. This model can be interpreted as a simplification/approximation of this more general framework, where monitoring costs for households are too high ($= \infty$) and the returns follow a binomial process.

[14] Similar to Holmstrom and Tirole [1997], in an alternative interpretation, households deposits their money with the bank which in turn lend those funds to firms.

[15] Entrepreneurs can potentially save at the risk-free rate by lending to the unconstrained households, i.e. $B_{i,t}$ is allowed to be negative. Most of the time this won't happen.

I refer to R^B_{t+1} as the *loan rate* and $\bar{R}^B_{t+1}$ as the *expected loan rate*. The loan rate is an endogenous equilibrium object that clears the market for bank financing and potentially the expected loan rate can be (and is in most of the analysis) different from the riskless return R. I refer to the spread $\bar{R}^B_{t+1} - R$ as the loan premium. Importantly, the loan premium is net of default probabilities, so it represents the price that banks charge for holding risk (*price of risk-sharing*). In the model, all agents are risk neutral so variations in the loan premium are related, instead, to the tightness of constraints.

3.2.3 Capital investment and the down-payment required

How much capital can firms invest for given net worth $n^E_{i,t}$? Borrowing constraints imply that the capital investment is constrained by the net worth of the firm. From (6), (7) and (8) we get

$$k_{i,t} \leq \frac{1}{\theta_t} n^E_{i,t}, \tag{9}$$

where the variable θ_t represents the minimum down-payment required by the entrepreneur to purchase a unit of capital.[16] We can write $\theta_t = \theta^H \theta^B_t t(\theta^H, \theta^B_t)$ with $\theta^H = 1 - (1-\kappa)/R$, $\theta^B_t = 1 - \bar{\phi} \mathrm{E_t} R^k_{t+1} / \left(\bar{p} R^B_{t+1}\right)$ and $t\left(\theta^H, \theta^B_t\right) = 1 - \left(1/\theta^H - 1\right)\left(1/\theta^B_t\right)\bar{\phi}\left(\bar{R}^K - R\right)/\left(\bar{p} R^B_{t+1}\right)$.

If we switch off the lending channel from banks ($\bar{\phi} = 0$), then $\theta_t = \theta^H$ which corresponds to the required down-payment in Kiyotaki and Moore [1997]. If we instead switch off the direct lending channel from households ($\kappa = 1$), then $\theta_t = \theta^B_t$ which depends on the loan rate R^B_{t+1}. When we have both lending channels present, the down-payment required is even lower $\theta_t < \theta^H \theta^B_t$. This is because there is complementarity between the two lending channels.[17]

The maximum amount of funds firms can borrow and invest in capital is determined by their net worth $n^E_{i,t}$ and the down-payment required θ_t, which depends on the endogenous loan rate R^B_{t+1}.

3.2.4 Net worth dynamics and the firms' problem

Entrepreneurs are financially constrained and, absent some motive for paying dividends, may find it optimal to retain earnings and accumulate net worth to the point where constraints no longer matter. In order to limit entrepreneurs' ability to save to overcome financial constraints, I assume that firms exit with probability τ. In case of exit, entrepreneurs consume all their wealth at this time.[18]

[16] θ_t represents the margin required by creditors and $1/\theta_t$ the maximum leverage

[17] If an entrepreneur only uses direct lending, she would invest $1/\theta^H$. If given that size of the balance sheet she then borrows the maximum she can from banks, then she would invest $\left(1/\theta^H\right)\left(1/\theta^B\right)$. But, given this new size of balance sheet she can borrow even more from households, etc. This multiplier effect is captured by $t(\theta^H, \theta^B)$.

[18] Equivalently, we can interpret this as entrepreneurs become households upon exiting.

Surviving entrepreneurs start the period with net worth $n^E_{i,t}$, they borrow funds from households and banks through debt and loans and invest in capital $k_{i,t}$, implying the following law of motion for net worth

$$n^E_{i,t+1} = R^k_{i,t+1}k_{i,t} - R^B_{i,t+1}L_{i,t} - RB_{i,t} \tag{10}$$

Therefore, surviving entrepreneurs choose $(B_{i,t}, L_{i,t})$ to maximize their expected wealth upon exiting, given the law of motion of net worth and borrowing constraints and taking as given the loan rate R^B_{t+1}. We write entrepreneurs' problem as follows

$$\max_{B_{i,t},L_{i,t}} \sum_t^{\infty} (1-\tau)^{t-1}\tau \mathrm{E}n^E_{i,t}$$

s.t. (10), (6),(7) and (8).

3.3 Banks

There are a continuum of agents who run financial firms, I refer to these agents as banks. Banks are special because their monitoring advantage allows them to share firms' idiosyncratic risks. Banks are born with this monitoring technology and I assume there is zero cost to use it. Thus, banks always monitor their loans. As explained above, monitoring outcomes (firms' reports to banks) are privately observed by banks and so, similar to Krasa and Villamil [1992], when risks cannot be perfectly diversified (aggregate risks) the problem of monitoring the monitor arises. Because households cannot monitor, neither firms nor banks, thus banks can only borrow from households through riskless debt (i.e. deposits) which I denote by D_t.

Banks are indexed $j \in [1,2]$. Bank j starts period t with net worth $n^B_{j,t}$, borrows from households by issuing riskless deposits $D_{j,t}$ and lends to firms. Each bank lends to multiple firms. I denote $L_{j,t}$ as the total amount of loans of bank j, thus $L_{j,t} = \int_i L_{i,j,t}di$ where $L_{i,j,t}$ is the specific bank j loan to firm i

$$L_{j,t} = n^B_{jt} + D_{j,t}. \tag{11}$$

Thus, banks' assets consist of loans to firms. I denote $\tilde{R}^B_{t+1}$ as the total return on banks' assets (return on their portfolio of loans), then

$$\tilde{R}^B_{t+1} = \frac{\int_i R^B_{i,t+1}L_{i,j,t}di}{L_{j,t}}.$$

Banks' debt must be riskless, i.e. deposits must be secured by the lowest realization of return on assets, which implies the borrowing constraint

$$RD_{j,t} \leq \tilde{R}^B_{L,t+1}L_{j,t}, \tag{12}$$

where $\tilde{R}^B_{L,t+1}$ denotes the lowest possible return on banks' assets $\tilde{R}^B_{t+1}$.

By diversifying risks, banks increase $\tilde{R}^B_{L,t+1}$ and relax their constraints. Since banks have zero cost to pool over different firms, in equilibrium they perfectly diversify idiosyncratic risks, thus

$$\tilde{R}^B_{t+1} = \int_i R^B_{i,t+1} = p_{t+1} R^B_{t+1} \tag{13}$$

and

$$\tilde{R}^B_{L,t+1} = p_L R^B_{t+1}.$$

While banks can diversify out any idiosyncratic risk from their portfolio, they hold aggregate risks in their balance sheets: while the loan rate R^B_{t+1} is known at t, the fraction of firms that default on their loans p_{t+1} is stochastic and involves an aggregate risk that banks cannot diversify away. Thus, they require net worth to repay their debt in case of negative aggregate shocks. Therefore, the size of their balance sheets depends on their net worth, (11) and (12) lead to

$$L_{j,t} \leq \frac{1}{\tilde{\theta}_t} n^B_{j,t}, \tag{14}$$

with $\tilde{\theta}_t = 1 - \tilde{R}^B_{L,t+1}/R$. Condition (14) says that in order to fund one unit of loans, banks need to at least finance the down-payment $\tilde{\theta}_t$ from their own net worth. This down-payment (or margin) is directly related to the return on the bank's portfolio of loans and more importantly to its lowest possible realization $\tilde{R}^B_{L,t+1}$, which we can think of as representing the collateral value of their assets.

3.3.1 Net worth dynamics and the banks' problem

Like entrepreneurs, banks exit with probability τ and consume all of their wealth.[19] Banks keep accumulating wealth until they exit. Thus, the net worth of a surviving bank j, who provides loans $L_{j,t}$ and issues deposits $D_{j,t}$, evolves as

$$n^B_{j,t+1} = \tilde{R}^B_{t+1} L_{j,t} - R D_{j,t}. \tag{15}$$

Each bank maximizes

$$\max_{D_{j,t}} (1-\tau)^{t-1} \tau E n^B_{j,t},$$

subject to (15), (12) and (11).

[19] To maintain a non-degenerate cross-sectional distribution of firms and banks, I assume that with exogenous probability τ^{entry}, households become firms/banks with random initial net worth with mean equal to the average net worth of firms/banks. In equilibrium, the state variable that matters is the aggregate net worth of borrowers

3.4 Intermediation: diversification and collateral

An important feature of the model is that intermediation is key in the provision of risk-free securities that can be used as collateral, which helps allocate resources to the most productive sector. Banks' ability to monitor allows them to share part of the projects' risks with firms, through risky loans. By lending to multiple firms, banks pool idiosyncratic risks and use their diversified portfolio as collateral to raise additional funds from households, funds that in turn banks lend to firms. While entrepreneurs are only able to borrow against the part of their assets that is not affected by idiosyncratic shocks, that is $(1-\kappa)(1-\delta)$, banks can borrow against the fraction p_L of assets that on average are not affected by the idiosyncratic shock. In that sense, banks' ability to share and diversify risks allows them to provide "new collateral".

I refer as intermediation to this process by which banks share risks and create collateral, allowing more funds to flow from households to firms, through banks, and improving the allocation of capital. Importantly, the size of intermediation (how much funds banks can intermediate from households to firms) depends on the amount of loans banks hold on their balance sheets. On the liability side, because banks hold the aggregate risk (associated with p_{t+1}), the size of their debt and so the size of their of their balance sheets depends on banks' net worth.[20] On the assets side, banks intermediate only over the size of the loans firms can borrow, which is constrained by firms' net worth given the maximum promised payment represented by $\bar{\phi}$.

Therefore, in equilibrium the size of intermediation and banks' balance sheets fluctuates depending on the endogenous evolutions of banks' and firms' net worths.

3.5 Equilibrium

A competitive equilibrium consists of sequences for the aggregate allocation of capital K_t and loan rates $\left\{R^B_{t+1}\right\}_0^\infty$, and allocations for each firm $\left\{n^E_{i,t}, k_{i,t}, B_{i,t}, L_{i,t}\right\}_0^\infty$ and for each bank $\left\{n^B_{j,t}, L_{j,t}, D_{j,t}\right\}_0^\infty$, such that[21]:

1. taking prices as given, the allocations solve the optimization problems of firms and banks,

2. the market for bank lending clears

$$L_t = \int_j L_{j,t} dj = \int_i L_{i,t} di$$

The deep-pocket feature of households with an exogenous discount rate implies the supply of

[20] Market incompleteness for aggregate risks is a common feature in the literature. Di Tella [2012] suggests uncertainty shocks as a possible explanation for the excessive exposure of intermediaries to aggregate risks, even under an available market to trade such aggregate risk.

[21] The sequence of allocations are adapted to the sequence of shocks $\{p_s\}_0^t$

riskless debt or deposits is residual: it is perfectly elastic, and adjusts to satisfy any demand that provides the riskless return R.

4 Solving the equilibrium

4.1 Benchmark without borrowing constraints

Without borrowing constraints balance sheets do not play any role. Entrepreneurs, banks and households frictionlessly share risks and, given they are risk neutral, arbitrage away any spreads $\mathrm{E_t} R^k_{t+1} = \bar{R}^B_{t+1} = R$. Decreasing returns in aggregate capital implies there is an efficient allocation of capital K^{FB} which, using (5), solves

$$\mathrm{E_t} R^k_{t+1} = \bar{p} A_{t+1} + (\bar{p}\kappa + (1-\kappa))(1-\delta) \quad = \quad R. \tag{16}$$

Under the assumption of $A_t(K_t) = \alpha K_t^{\alpha-1}$,

$$K^{FB}_t \quad = \quad \left[\frac{\alpha\bar{p}}{R - (\bar{p}\kappa + (1-\kappa))(1-\delta)}\right]^{\frac{1}{1-\alpha}}$$

Given that shocks p_{t+1} are i.i.d., the economy instantaneously arrives to its efficient level of capital and remains there at all periods. The net worths of both entrepreneurs and banks grow at rate $R(1-c)$.

4.2 Linear policies, aggregation, and the state space

In a competitive equilibrium, entrepreneurs and banks take R^B_{t+1} as given, and face dynamic problems with linear constraints in net worth and constant returns on their investments, which implies optimal policies are linear in net worth, $(k_{i,t} = \hat{k}_t n^E_{i,t}, B_{i,t} = \hat{B}_t n^E_{i,t}, L_{i,t} = \hat{L}_t n^E_{i,t})$ and $(L^B_{j,t} = \hat{L}^B_t n^B_{j,t}, D_{j,t} = \hat{D}_t n^B_{i,t})$. This allows aggregation across entrepreneurs and across banks and simplification of the state space: we need to keep track of the aggregate net worths, separately, of firms and banks, but the distribution across each kind of agent is not important. We can represent the economy with a representative firm, a representative bank, and a representative household, which behave competitively and are subject to the described constraints.

Also we do not need to keep track of the net worth of households (implicitly we are assuming it is sufficiently large), R is exogenous, nor the level of capital. As a consequence, the state space of the economy is summarized by the aggregate net worths of firms and banks $s_t = \left(N^E_t, N^B_t\right)$. Equivalently, the state space (N^E, N^B) can be represented by the total net worth in both sectors and the share of banks' net worth $(N = N^E + N^B, \beta = N^B/N)$.

The literature has mainly focused on models where the aggregate net worth of constrained borrowers, N, is the relevant variable which determines the allocation of capital and economic dynamics, while the shares of net worth between different borrowers β has not played any role. In the models of Bernanke and Gertler [1989], Kiyotaki and Moore [1997], Bernanke et al. [1999], Jermann and Quadrini [2012] or Christiano, Motto, and Rostagno [2014], constrained entrepreneurs are the borrowers and their net worth is the relevant variable; intermediaries/banks are merely a veil or not present at all, thus these models are represented by the particular case when $N^B = 0$ or $\beta = 0$[22]. Instead, in the models of He and Krishnamurthy [2013] or Adrian and Boyarchenko [2012], the borrowers are financial experts, intermediaries or banks, which directly manage the productive assets of the economy. Their net worth is the relevant state variable and models are calibrated to net worth or leverage data of the financial sector: firms are not present in these cases thus they correspond to the case when $N^E = 0$ or $\beta = 1$.[23] Finally, Gertler and Kiyotaki [2010], Gertler and Karadi [2011], Brunnermeier and Sannikov [2012] model explicitly both a financial sector and a productive sector, but there is a frictionless relationship between the agents in each sector, which implies that the relevant variable is the sum of the aggregate net worth $N = N^E + N^B$.

In my model, banks have frictions on both sides of their balance sheets. Banks and firms may have different borrowing and investment strategies and so the conditions of their balance sheets may play different roles and differently affect output dynamics. Consequently, the dynamics of the economy additionally depend on the state of the share of banks' net worth β.

I divide the state space into different regions according to which borrowing constraints are binding and which are not. In each of these regions banks' and firms' net worths play different roles and each region features very different economic dynamics. As depicted in Figure 3, the economy is characterized by three constraints: firms-households, firms-banks, and banks-households. The partition is characterized by four different regions,[24] as depicted in Figure 4: (i) the region where both banks and firms borrow up to the limit of their constraints, which I call the regime of normal booms and recessions; (ii) the region where firms borrow below the maximum from banks, which corresponds to financial crises; (iii) the region where banks borrow below the maximum from households; (iv) the region where banks and firms borrow below the limit from households, which corresponds to the unconstrained regime.

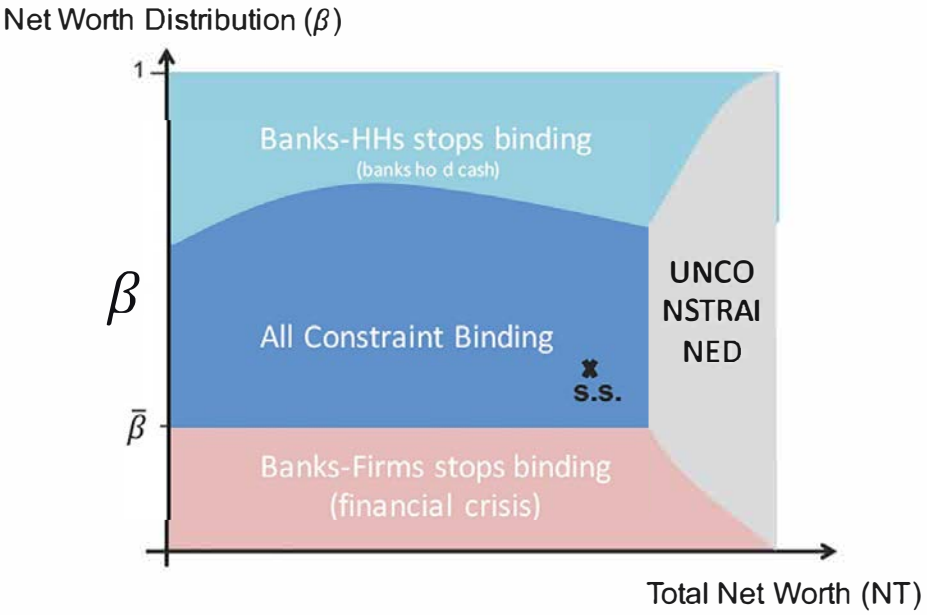
This partition is endogenous: agents optimally decide to borrow up to the limit of their constraints or not. Thus, the boundaries depend on the agents policies and vice versa. I focus on the first two regimes, which I consider the most interesting and are also the ones used in

[22] There are many different details among the models, but the balance sheet mechanism is basically the same.

[23] In my model, this also requires that there is no friction between banks and firms ($\bar{\phi} = 1$), so that banks can continue lending even if firms' net worth is zero.

[24] Note that there are potentially eight possible combinations, depending on which constraints are binding or not.

Figure 4: State Space Partition



the estimation. Thus, throughout the paper, I assume that both banks and firms borrow from households up to the limits of their constraints.

I proceed first by describing the dynamics of the model during normal times (taking boundaries as given), consecutively I explain how to find these boundaries, and then I explain the dynamics during the financial crisis regime. Before that, I characterize the "risky" steady state, which is the state in which the economy converges if shocks realize at the expected value.

4.3 Risky steady state

The risky steady state is defined as the limit where the economy converges if agents expect future risk but shocks are realized at their expected values, i.e. $p_t = \bar{p}, \forall t$. I solve for the risky steady state under the assumption/claim[25] that at this limit all constraints bind.

The risky steady state allocation of capital K^{ss} (under binding constraints):

$$\bar{p}A_{t+1}(K_t^{ss}) = \frac{[R-(\bar{p}\kappa+(1-\kappa))(1-\delta)]+p_L\bar{\phi}\kappa(1-\delta)(1-R(1-c))}{R(1-c)} \tag{17}$$

The first term in the numerator of (17) corresponds to the unconstrained case described in (16). We can see that $R(1-c)<1$ implies $K^{ss}<K^{FB}$, firms and banks die fast enough and do not accumulate enough net worth to get away the constraints. Throughout the paper, I

[25]This claim depends on the parameter values, thus the assumption is that the relevant calibration satisfies this property.

maintain this assumption.

Assumption 1. $R(1-c)<1$

4.4 Normal booms and recessions

In this section, I describe the dynamics where both banks and firms borrow up to the limit of their constraints (all constraints bind). By assumption, the risky steady state belongs to this region. For small shocks the economy lives inside this region and this is why I named it times of normal booms and recessions.

In this region, the investment in capital is characterized by the following proposition.

Proposition 1. *When firms borrow up to the maximum from both households and banks, and banks borrow up to the maximum from households, the equilibrium investment in capital is*

$$K_t = \bar{\lambda}\left(N_t^E + N_t^B\right) = \bar{\lambda} N_t \tag{18}$$

with

$$\bar{\lambda} = \left(1 - \frac{1}{R}\left(p_L \bar{\phi}\kappa + (1-\kappa)\right)(1-\delta)\right), \tag{19}$$

and the evolution of total net worth follows

$$N_{t+1} = (1-\tau)\left(p_{t+1} R_{t+1}^k K_t - R(B_t + D_t)\right) \tag{20}$$

Proposition (1) implies that the sum of net worth N_t determines capital investment and output. Firms' and banks' net worths have the same effect on investment and output. In this sense, both net worths play the same role in the dynamics of the economy. Notice that both banks and firms are important and the dynamics would be different if any of them were not present. The key is that both have the same investment incentives[26], and what matters is how much funds they can together borrow from households and invest in the productive technology. Equation (18) shows that the total funds they invest depend on their total net worth and the multiplier $\bar{\lambda}$. Equation (19) shows that the multiplier depends on the value of collateral that both firms and banks are able to provide to households. Firms provide $(1-\kappa)(1-\delta)$ units of collateral for every unit of capital. Banks, by diversifying risks, provide the additional collateral $p_L\kappa(1-\delta)$ per unit of capital but only of the part of assets that they are able to intermediate, which is limited by $\bar{\phi}$.

The transmission mechanism stressed by Bernanke and Gertler [1989] or Kiyotaki and Moore [1997], by which the effects of shocks persist and amplify, works through the aggregate balance

[26] In this region both borrow up to the maximum of their constraints.

sheet of firms and banks. A fall in total net worth N_t implies a reduction in capital K_t and affects future net worth N_{t+1} (see (20)). For instance, this mechanism propagates the aggregate shocks to p_{t+1} as follows: because both banks and firms are levered, a temporary negative shock to p_{t+1} reduces total net worth and investment. The effects persist because the fall in future capital makes cash-flows, future net worths, and future investment lower than it would otherwise be. This channel, the so-called financial accelerator, induces persistence through balance sheets[27]. However, balance sheet cycles are driven by total net worth, N_t, but not by the distribution of net worth in each sector, β_t. Importantly, the transmission mechanism depends on the parameter $\bar{\phi}$, which represents how much intermediation is in the economy. A more financially developed sector (high $\bar{\phi}$) will have different dynamics than a less developed one (low $\bar{\phi}$).

4.4.1 Market of bank loans

While the share of net worth β_t does not affect the investment on capital, it does affect how the returns on capital are distributed between each sector. The loan rate R^B_{t+1} clears the market of bank loans and, as a result, the loan rate depends separately on both the net worth of banks and firms. In particular, the equilibrium loan rate can be described by

$$R^B_{t+1} = r^B\left(N^E_t, N^B_t\right), \tag{21}$$

where the function r^B is increasing in N^E_t and decreasing in N^B_t.

The intuition behind equation (21) is the following. On the one hand, the aggregate demand on bank loans depends on firms' net worth: an increase in firms' net worth slackens firms' constraints and reduces firms' shadow costs of borrowing from both households and banks. This increases the demand for bank loans and pushes the loan rate upwards. On the other hand, the aggregate supply of bank loans depends on banks' net worth: an increase in banks' net worth slackens banks' constraints and reduces banks' shadow costs of borrowing from households. This increases the bank demand for deposits, increases the supply of bank loans to firms, and pushes the loan rate downwards.

4.4.2 Effects of a redistribution between borrowers

In this region of normal booms and recessions, a redistribution of funds from banks to firms, or vice versa, does not have any effect on investment or output. On the one hand, the increase in firms' net worth slackens firms' constraints and lowers borrowing costs (shadow value) for

[27] In the same way, a dynamic amplification channel, as in Kiyotaki and Moore [1997], appears if we include effects through asset prices, for example by introducing capital adjustment costs. It is important to notice that the transmission of the dynamic channel would also only work through the total net worth N_t.

firms. On the other hand, the drop in banks' net worth tightens banks' constraints and raises borrowing costs (shadow values) for banks. Banks pass on these costs to firms by increasing loan rates, creating no effect on total borrowing or investment.

4.4.3 Representative borrower representation

For small shocks the economy lives inside this region where all constraints always bind. Therefore, a corollary of Proposition (1) follows: we can merge banks and firms in one sector represented by a joint "entrepreneur-bank" agent, a representative borrower, which is constrained in borrowing from households. This borrower owns the total net worth N_t and requires a down-payment $1/\bar{\lambda}$ to purchase a unit of capital (the rest, $1-1/\bar{\lambda}$, is borrowed from households). The distribution of net worth between firms and banks only affects how returns or profits are shared between these two agents (affects the bank loan rate), but it doesn't affect capital or output dynamics.

As commented in section (4.2), the literature mainly has abstracted from one of these sectors or has considered a frictionless relation. Proposition (1) implies that if the focus is on output dynamics and capital allocations, even when there is a frictional relation but constraints bind, such an abstraction gives a good approximation of the economic dynamics close to the steady state. In addition, models should be calibrated to aggregate net worth (and average leverage of the financial and non-financial sectors) and leverage constraints should be interpreted as a combined friction from both sectors.[28]

5 Financial crises

The financial crises regime is defined as the region of the state space where firms stop borrowing from banks up to the maximum of the constraint. This happens when the net worth of banks is critically low and the loan rates so high such that firms find bank loans unattractive and cut their borrowing.

5.1 Firms' optimal borrowing decision

The decision of firms to borrow from banks up to the maximum of the constraint depends on the loan rate R^B_{t+1}. The marginal benefit for a firm that borrows an additional unit from banks depends on the project's return on capital, while the marginal cost depends on the loan rate. The following proposition states that there is an upper bound such that for loan rates beyond that threshold firms stop borrowing from banks.

[28] For example, Gertler and Kiyotaki [2010] and Gertler, Kiyotaki, and Queralto [2012] calibrate parameters to match an average leverage ratio across different sectors (financial and non-financial).

Lemma 1. *There exists a bound $R^B_{t+1,max}$ such that firms' loan demand is characterized by*

$$L_{i,t} = \begin{cases} \bar{L}_{i,t} & \text{if } R^B_{t+1} < R^B_{t+1,max} \\ \in \left[0; \bar{L}_{i,t}\right] & \text{if } R^B_{t+1} = R^B_{t+1,max} \\ 0 & \text{if } R^B_{t+1} > R^B_{t+1,max} \end{cases},$$

where $\bar{L}_{i,t} = \bar{\phi}\kappa(1-\delta)k_{i,t}/R^B_{t+1}$ represents the maximum amount allowed by the (IC) constraint. The bound $R^B_{t+1,max}$ is a function of the return on capital R^k_t.

5.2 The partition

Recall that the equilibrium loan rate clears the market for bank loans and depends negatively on the share of banks' net worth β_t. Thus, the maximum loan rate $R^B_{t+1,max}$ is associated with a threshold $\bar{\beta}$, such that when the share of banks' net worth is below this threshold, the economy moves into the financial crises regime.

Lemma 2. *There exists a bound $\bar{\beta}$ such that for $\beta_t < \bar{\beta}$ the economy enters into the financial crises regime. In this regime, the equilibrium loan rate is at the maximum $R^B_{t+1} = R^B_{t+1,max}$ and firms borrow from banks below the limit of the constraint. The threshold $\bar{\beta}$ is a constant.*

Lemma 2 defines the partition of the state space that defines the financial crises regime.

Definition 1. The region of the state space that is associated with *financial crises* is defined as the set $\left\{(N_t, \beta_t) : \beta < \bar{\beta}\right\}$.

5.3 Characterization of equilibrium

During financial crises, capital investment depends separately on firms' and banks' net worth. Banks intermediate, diversify and create collateral, over the size of the loans that they hold on their balance sheets. The reduction in lending implied by the high spreads during financial crises impacts the ability of banks to do intermediation. This generates an additional effect on investment and output. The following proposition characterizes the investment of capital in this regime.

Proposition 2. *During financial crises, when both firms and banks borrow from households up to the maximum of their constraints but firms borrow below the maximum from banks (i.e. $\beta_t < \bar{\beta}$), the equilibrium capital investment is*

$$K_t = \lambda_t N_t, \tag{22}$$

with

$$\lambda_t = 1 - \frac{1}{R}\left(p_L \phi_t \kappa + (1-\kappa)\right)(1-\delta), \tag{23}$$

with

$$\phi_t = \frac{R^B_{t+1,max} L_t}{\kappa(1-\delta)K_t}, \tag{24}$$

where $\phi_t = \bar{\phi} L_t / \bar{L}_t < \bar{\phi}$ *represents the aggregate demand for bank loans relative to the maximum allowed by the constraint. In addition,* $\phi_t = \phi(\beta_t)$ *depends positively on* β_t *and* $\phi(\bar{\beta}) = \bar{\phi}$.

Proposition 2 implies that, in the financial crisis regime, the dynamics of capital investment and output depend on the total net worth N_t and on the the share of banks' net worth β_t. The multiplier λ_t depends on β_t through the relative size of the loans that banks hold on their balance sheets which is captured by ϕ_t, representing the size of intermediation: the amount of collateral that banks create is $p_L \phi_t \kappa$ per unit of undepreciated capital (see 23). In this regime, the net worth of banks is special and has a stronger impact on investment and output than firms' net worth: an increase in banks' or firms' net worth increases the level of capital and so the amount of collateral that can be used to borrow from households (captured in (22) by an increase in N_t), in addition, an increase in banks' net worth increases intermediation and so the fraction of capital that can be used as collateral (captured in (22) by a change of λ_t).

5.4 Intermediation mechanism

A new mechanism appears during financial crises. In this regime, shocks to banks' net worth impact the intermediation process: a drop in banks' net worth tightens banks' constraints, raises banks' borrowing costs (shadow costs) for banks and contracts the loan supply. This puts upward pressure on the loan rate, which is already at the maximum firms are willing to accept, i.e. R^B_{max}. Thus, banks cannot pass the increase in borrowing costs on to firms. This implies a quantity adjustment and a contraction in the amount of bank loans. Banks have to reduce their balance sheets: sell their assets and reduce their extended loans. This lowers intermediation and reduces the collateral banks use to borrow from households, which in turn tightens banks' constraints further and raises banks' borrowing costs even more. This feedback loop, generated through the *intermediation mechanism*, leads to a collapse of banks' balance sheets and severely affects investment and output. Note that the reduction in the flow of funds also occurs between households and firms because the reduction in capital decreases the collateral that secures corporate bonds.

This intermediation mechanism appears in addition to the standard financial accelerator of Bernanke and Gertler [1989]. The mechanism appears even when total net worth remains constant. In this regime, a redistribution of funds from banks to firms worsens the recession. This

may be counterintuitive as firms are the ultimate borrowers that are financially constrained, but notice that a dollar in the hands of banks allows more dollars to be invested in firms because banks can lever and lend more (by diversifying) than what firms directly could.

There is a complementarity between the standard financial accelerator mechanism and this new "intermediation mechanism". The law of motion of total net worth is described by

$$N_{t+1} = (1-\tau)\left(R_t^k K_t - R(B_t + D_t)\right).$$

Therefore, the intermediation mechanism triggers the financial accelerator: shocks to banks' net worth and so to the distribution β_t, even under constant N_t, affect capital investment and output. In turn, this affects cash-flows and future net worths of both banks and firms N_{t+1} (future balance sheets). Therefore, even a temporary shock to β_t triggers the credit cycle of Bernanke and Gertler [1989] in the next period

$$\beta_t \downarrow \Rightarrow \phi_t \downarrow \Rightarrow K_t \downarrow \Rightarrow Y_{t+1} \downarrow \Rightarrow N_{t+1} \downarrow \Rightarrow \text{"BG mechanism"}.$$

In addition, the dynamic of the share of banks' net worth β_{t+1} is endogenous. The law of motion of banks' net worth is described by

$$N_{t+1}^B = (1-\tau)\left(R_{max}^B L_t - RD_t\right).$$

In this regime, banks severely contract their balance sheets, and this reduction of loans lowers their exposure to the returns on firms' projects, which in expectation delivers high excess returns. This implies that any shock, even temporary, that affects the share of banks' net worth and brings the economy to the financial crisis regime has additional persistent effects through the intermediation mechanism. It takes time to exit the financial crisis region.

$$\beta_t \downarrow \Rightarrow \phi_t \downarrow \Rightarrow \beta_{t+1} \downarrow \text{ (persistence)}.$$

In the next section, I include different shocks that move the economy across regions, and I use the model to explain the observed dynamics of output and net worth.

6 Quantitative Exercise

In this section, I quantify the importance of the new dynamics that appear during the financial crisis regime. I exploit the non-linear relationship between banks' net worth and economic activity using the structure of the model. I estimate the underlying parameters to quantify the importance of the intermediation mechanism highlighted in the previous sections. In particular,

the quantitative exercise provides answers to the following questions: i) in which periods was the mechanism activated (i.e. the economy entered into the financial crisis regime)? ii) how much of the variation in output and net worth is explained by this endogenous mechanism? and iii) how would model forecasts change when we take into account the intermediation mechanism?

6.1 Extended model

I now augment the model with two shocks: an aggregate shock to the marginal productivity of capital Z_t^Y and an exogenous shock to banks' net worth Z_t^{NB}. The purpose of Z_t^{NB} is to capture the losses in banks' investments that are not in the model. For instance, the Great Recession was triggered by the housing crisis, at the onset of the crisis a big share of banks' assets were invested in securities related to the housing market, which severely affected banks' balance sheets. My model abstracts from mortgages and the housing sector, so I capture such variations in the data with Z_t^{NB}. To be consistent with the literature, I generalize aggregate technology shocks Z_t^Y to feature persistence, and I model them as shocks to the productivity of capital, instead of shocks to the probability p_t as described in section 3. Both exogenous shocks, Z_t^Y and Z_t^{NB}, follow AR(1) processes:

$$\begin{aligned} Z_t^Y &= \rho^Y Z_{t-1}^Y + \sigma^Y \epsilon_t^Y, \\ Z_t^{NB} &= \rho^{NB} Z_{t-1}^{NB} + \sigma^{NB} \epsilon_t^{NB}, \end{aligned}$$

where ϵ_t^K and ϵ_t^{NB} are non-correlated and normally distributed innovations.

I assume the economy always lives in the two regions described in the previous sections: around-steady-state (normal times) and financial crises, i.e. both firms and banks are always constrained in their borrowing from households. This is done for simplicity, as the non-linear features of the model make the estimation slow/heavy. For these two regions I am able to derive closed form solutions and so the estimation is considerably faster when looking just at these two. The point of this section is to provide the quantitative implications of the mechanism that appears during financial crises, so focusing on these two regions is enough.

The model I estimate is described in full by the following equations: aggregate output

$$Y_t = \exp\left(Z_t^Y\right) K_{t-1}^{\alpha};$$

aggregate capital K_t invested by entrepreneurs

$$K_t = \lambda_t\left(N_t\right);$$

the multiplier λ_t that determines leverage of banks and firms (the merged borrowing sector)

$$\lambda_t = \begin{cases} \bar{\lambda} & \text{if } N_t^B/N_t > \bar{\beta} \\ \lambda(N_t^B/N_t) & \text{if } N_t^B/N_t \leq \bar{\beta} \end{cases},$$

where $\bar{\lambda}$ and $\lambda(N_t^B/N_t)$ follow from Propositions 1 and 2. The laws of motion of total net worth and the net worth of banks

$$\begin{aligned} N_{t+1} &= (1-\tau)\left(\left(R_{t+1}^K - R\right)\lambda_t + R\right)N_t, \\ N_{t+1}^B &= (1-\tau)\left(\tilde{R}_{t+1}^B L_t - RD_t\right)\exp\left(Z_{t+1}^{NB}\right), \end{aligned}$$

firms' credit from households B_t and from banks L_t, and banks' credit from households (deposits) D_t

$$\begin{aligned} B_t &= \frac{(1-\kappa)(1-\delta)}{R}K_t, \\ L_t &= K_t - B_t - N_t^E, \\ D_t &= L_t - N_t^B, \end{aligned}$$

finally, the return on firm assets and the return on bank loans

$$\begin{aligned} R_{t+1}^K &= \exp\left(Z_{t+1}^Y\right)K_t^{\alpha-1} + (1-\delta), \\ \tilde{R}_{t+1}^B &= \exp\left(Z_{t+1}^Y\right)R_{t+1}^B, \end{aligned}$$

where $R_{t+1}^B = r^B(N_t^B, N_t)$ follows from Equation 21.

6.2 Estimation

I estimate the model to fit the time series variation of output (Y) and banks' net worth (N^B). I measure output using quarterly U.S. real GDP and banks' net worth as the real aggregate market value of equity of U.S. commercial banks (from the U.S. Financial Accounts). Prior to analysis, I remove linear trends and demean each series. The estimation method uses information from the correlations and autocorrelation functions, but it doesn't make use of information on average ratios (steady states). The sample covers the period from 1980:Q1 to 2015:Q4.

Table 1: Estimated Parameters

Parameter	Description	Value
α	technology parameter	0.49
τ	exit probability of banks and firms	0.041
κ	fraction of capital hit by idiosyncratic shocks	0.15
$\bar{\phi}$	maximum share of firms' returns held by banks	0.17
ρ^Y	autocorrelation, technology shock	0.59
ρ^{NB}	autocorrelation, net worth shock	0.67
σ^Y	standard deviation, technology shock	0.007
σ^{NB}	standard deviation, net worth shock	0.14

The system can be represented by

$$\begin{aligned} \mathbf{obs_t} &= f(\mathbf{s_t}, \mathbf{Z_t}; \theta) \\ \begin{pmatrix} \mathbf{s_t} \\ \mathbf{Z_t} \end{pmatrix} &= g\left(\mathbf{s_{t-1}}, \mathbf{Z_{t-1}}, \epsilon_t; \theta\right)' \end{aligned} \tag{25}$$

where $\mathbf{obs_t} = (Y_t, N_t^B)$ is the vector of observables, $\mathbf{s_t} = (N_t, \beta_t)$ represents the vector of endogenous states and $\mathbf{Z_t} = (Z_t^Y, Z_t^{NB})$ the vector of exogenous states. The functions f and g represent the system of equations in section 6.1. The vector of model parameters is denoted $\theta = \left(R, \delta, \alpha, c, \kappa, \bar{\phi}, \rho^Y, \rho^{NB}, \sigma^Y, \sigma^{NB}\right)$.

I fix the risk-free rate and the depreciation rate to standard values: $R = 1.01$ and $\delta = 0.03$ and estimate four model parameters $(\alpha, c, \kappa, \bar{\phi})$ and four parameters related to the shock processes $(\rho^Y, \rho^{NB}, \sigma^Y, \sigma^{NB})$.

I use a maximum likelihood approach to estimate the model. This is a non-linear dynamic system with two exogenous states (Z^K, Z^{NB}) and two endogenous states (N^T, β) that are treated as latent variables. I construct the likelihood function following Fernández-Villaverde and Rubio-Ramírez [2007], which uses a particle filter as the updating procedure for latent variables.

Table 1 reports the estimated parameters. The estimated value for α is 0.49; α is the elasticity of output to capital investment which is driven by net worth in the model. This value is different than the standard capital share 0.35 as I do not use data on capital nor labor. The probability of exit τ is around 4.1%. The fraction of capital that is hit by the idiosyncratic shock κ is estimated to 0.15 and the risk-sharing parameter $\bar{\phi}$ is estimated as 0.17.

Table 2 reports the implied steady state properties for these parameters. The likelihood does not use information about these ratios, but I check that the parameters imply sensible values. The aggregate leverage ratio of the merged borrowing sector (firms and banks) is around

Table 2: Steady-State, Model vs Data

Variable	Model	Data
$\frac{K}{N^E+N^B}$	5.1	4 to 7 (literature range)
$\frac{N^B}{Y}$	0.07	0.09
$\frac{N^E}{N^E+N^B}$	0.95	0.91

5; Gertler, Kiyotaki, and Queralto [2012] calibrate their model to have an aggregate leverage ratio of 4, as an approximated mean across different sectors in the economy. The quarterly spread of the return of capital and the risk-free rate is around 2%, the ratio of banks' net worth to output and to total net worth are both around 10%, similar to the data.

7 Results

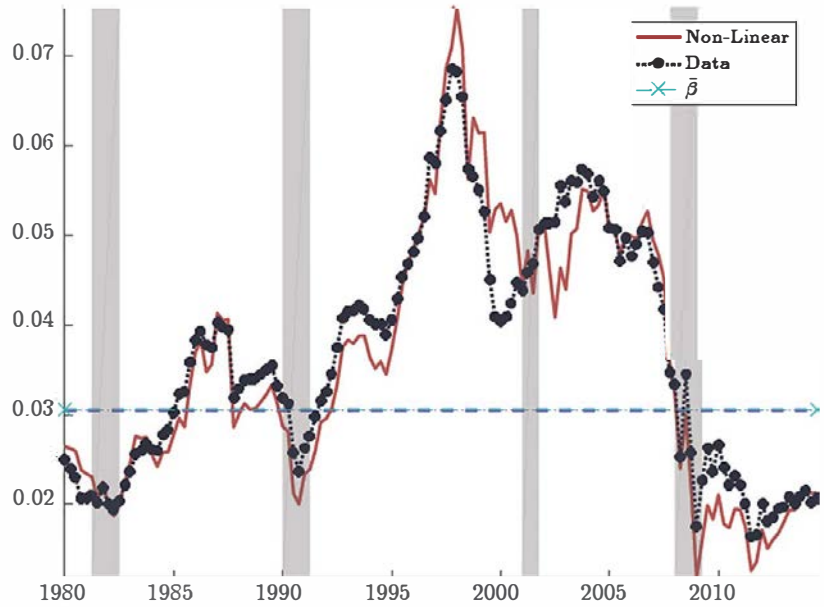
7.1 Financial crisis region

I start by presenting the periods where the U.S. entered into the financial crisis regime according to my model. Through the lens of my model a financial crisis is identified whenever the bank's net worth experiences a severe fall: $\beta < \bar{\beta}$. Figure 5 shows the estimated value of the latent distribution of net worth between banks and firms β, along with its data counterpart. The estimation procedure identifies the recessions of the 80s, early 90s, and 2008 as financial crises periods, as opposed to the 2001 crisis where the recession did not come along with a financial crisis. In addition, this methodology allow me to analyze the severity of financial recessions. The severity of the crisis is related to the size of the drop in banks' net worth. The recession of the 90s was a financial crisis but not as severe as in 2008. Finally, the net worth of banks has remained low for many periods after the last crisis, so the model indicates that the economy is still in the region featuring low risk-sharing and lending.

7.2 Importance of the intermediation mechanism

In order to assess the quantitative relevance of my mechanism I compare the implications of the non-linear model that features the intermediation mechanism to the model without the mechanism. The model without the mechanism is characterized by the equations described in 6.1 but fixing the risk-sharing parameter to $\phi = \bar{\phi}$ at all times, which implies a constant multiplier $\lambda = \lambda(\bar{\phi})$. Shutting off the mechanism can be interpreted as using the equilibrium equations during normal times (around-steady-state) to describe the whole dynamics. It is usual in the literature to consider the around-steady-state behavior to describe the whole dynamics

Figure 5: Estimated Banks' Net Worth Share β



Note: Banks' net worth share estimated using full dynamics along with data counterpart. Data is measured using the market value of equity of U.S. commercial banks and the market value of equity of non-financial corporate business from the Flow of Funds. The dashed horizontal line is the cutoff ($\bar{\beta}$) that defines the financial crisis regime. Shaded areas indicate NBER-dated recessions.

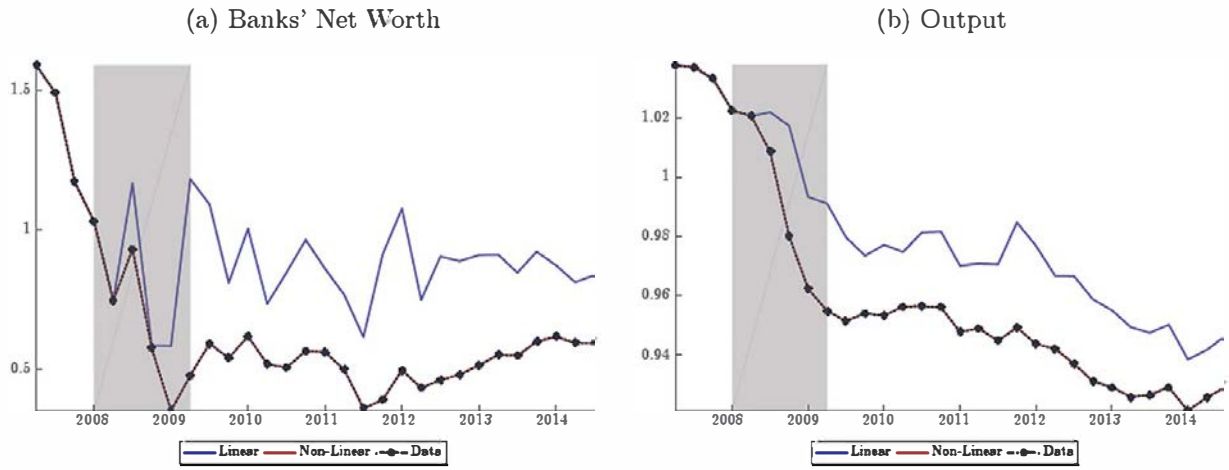
by linearizing, thus I refer to the dynamics without the intermediation mechanism as "linear dynamics" or "linear model". However, notice that the "linear dynamics" still come from the solution of the non-linear system (25), thus the only difference between both models is the intermediation mechanism which is captured by the change in the multiplier λ_t. References to the full model or "non-linear model" mean the dynamics implied by the full system (25).

Each model implies a different path for the estimated latent variables (endogenous states and exogenous shocks) and delivers different predicted dynamics for each series. In the first exercise, I fix the path of all latent variables to the ones recovered from the non-linear model and used them for both models (with and without the mechanism). In the subsequent exercises, I separately estimate the paths of latent variables for each case.

7.2.1 Contribution of the intermediation mechanism

I compare the full dynamics of the model relative to the counterfactual where the mechanism associated with financial crises is shut down. I use the latent variables and shocks (s_t, Z_t) of the full model, but I turn off the mechanism setting $\lambda_t = \lambda(\bar{\phi})$. This would be the counterfactual associated with the absence of the feedback loop generated by the intermediation mechanism. This statistical exercise identifies the variation in the data that can be explained solely by the

Figure 6: Contribution of Intermediation Mechanism



Note: Output and banks' net worth implied by the full model (red line) and model without the mechanism (blue line), along with the data counterpart (black dots). All series are shown as deviations from their linear trends. The model without the mechanism sets $\lambda_t = \lambda(\bar{\phi})$, but uses the estimated latent variables (endogenous states and exogenous shocks) from the full system.

endogenous mechanism.

Figure 6 displays the series of output and banks' net worth implied by the full dynamics (non-linear) and by the dynamics when the mechanism is turned off (linear) along with the data counterpart, for the periods after the recession of 2008. All series are shown as deviations from their linear trends. By construction, the non-linear model matches the data perfectly, as we are feeding in the shocks from the non-linear model. Using these shocks, we turn off the mechanism and find the counterfactual series in the absence of the mechanism. The differences between these series represents the variation in the data that is accounted by the endogenous intermediation mechanism. Table 3 reports the average size of the deviations of each series from its linear trend and the average difference as a percentage from the total observed drop. I conclude from this exercise that the mechanism explains on average 40% of the fall in output and 80% of the fall in net worth during the Great Recession.

7.2.2 Improvement of model fit and long horizon forecasts

In the following exercises, I estimate separately the latent variables (endogenous states and exogenous shocks) for each of the models with and without the mechanism. Both models use the same set of parameters θ from Table 1. First, I show that the model without the mechanism (linear) does not generate the multiplier effect during financial crises and instead estimates larger shocks to fit the data. Second, the model without the mechanism does not generate endogenous persistence. This implies a faster recovery for both net worth and output,

Table 3: Average Drop from Trend

Variable	Non-Linear (Data)	Linear	Mechanism Contribution (%Δ)
Banks' Net Worth (N^B)	35%	7%	82%
Output (Y)	4%	1.9%	44%

Note: Average deviation from linear trend for output and banks' net worth implied by the full model (matches observed data) and model without the mechanism (linear), from 2008:Q1 to 2015:Q1. The model without the mechanism sets $\lambda_t = \lambda(\bar{\phi})$, but uses the estimated latent variables (endogenous states and exogenous shocks) from the full system. %Δ: represents the average difference between the drop implied by the full model (observed drop) and the model without the mechanism as a percentage of the observed drop.

Table 4: Variance of Exogenous Shocks

	Linear	Non-Linear
s.d.$\left(Z^{NB}\right)$	0.44	0.32
s.d.$\left(Z^{Y}\right)$	0.018	0.017
Log-Likelihood	35.86	101.87

Note: This table reports the log-likelihood and the standard deviation of the two exogenous shocks for both models. Linear sets $\phi_t = \bar{\phi}$ in the system (25). Non-linear uses the full system (25). Each model separately estimates the latent variables (endogenous states and exogenous shocks).

which worsens long-horizon forecasts.

Improvement of goodness of fit. Table 4 reports the log-likelihood and the standard deviation of the two exogenous shocks for both models. The main message of this table is that the non-linear model does a better job in fitting the evolution of the time series of bank's net worth and real output. It also shows that the non-linear model requires a lower variance of the exogenous shocks to fit the evolution of the observables. I interpret these results as an improvement of the goodness of fit of the model and as a significant contribution of the intermediation mechanism.

Long horizon forecasting analysis. I have shown the ability of my model to fit the data used in the estimation. Now I conduct a different exercise in order to assess the prediction power of the intermediation mechanism for longer horizons. To do so, I compare the models' forecasting performance for short and long horizons, with and without the mechanism, for the periods after the Great Recession.

Let $\hat{obs}_{t/T_0}$ be the predicted value of the observables at time t conditioning on the information set up to time T_0 delivered by system (25), and let $\hat{\epsilon}_{T_0}(t) = log\left(\hat{obs}_{t/T_0}\right) - log\left(obs_t\right)$ be the

corresponding forecasting error.

Table 5 reports the root mean square of the forecasting error $RMSE(h) = \sqrt{\sum_{T_0} \left(\hat{\epsilon}_{T_0}(T_0+h)\right)^2}$ for different horizons. Sub-table 5a shows that banks net worth is a highly volatile series with an average short-term (1 quarter) RMSE of 15% and 10% for the linear and non-linear case, respectively. For longer horizons the model delivers a RMSE of around 30% for the linear dynamics but 17% for the non-linear case. The main message of this table is that the predictive power over banks' net worth improves at short- and longer-term horizons when non-linear dynamics are considered. Sub-table 5b shows the RMSE for output which is less volatile and presents smaller forecasting errors than banks' net worth. Still, the non-linear model delivers a lower RMSE than the linear model for longer horizons.

As an example, Figure 7 displays the actual values of the observables as a ratio of their linear trends (black dots) and their correspondent forecasts generated by both the linear (blue line) and the non-linear (red line) model. In Panel 7a, the model uses information up to the middle of the crisis, the fourth quarter of 2008, and delivers forecasts for the next four quarters. The nonlinear model outperforms the linear model in terms of forecasting power. It is more capable of replicating bot the level and the persistence of the fall that we see in the data.

The Great Recession features a slow recovery compared to other recessions; both output and banks' net worth dropped in 2008 and still remain below their trends. The intermediation mechanism implies a reduction of the net worth multiplier λ_t which leads to lower values of output per unit of net worth. Moreover, the mechanism implies a slow recovery for banks' net worth: when banks stop lending to firms, they stop sharing the cash-flows of projects and the expected returns on their net worth decreases. Hence, the non-linear dynamics induce persistence and improves the forecasts of net worth, which in turn helps in explaining the slow recovery of output relative to its trend.

7.3 External Validity: Disentangling Recessions

The model has also implications for the dynamics of the composition of credit and the cost of credit during different recessions. In this section, I study the model implied behavior of these variables for the recessions of the early 1990s, 2001, and 2008, and compare them to their aggregate data counterpart. It is important to remark that the data analyzed in this section was not used in the estimation process and thus it serves as external validation of the non-linear dynamics in the model.

7.3.1 Bank loans' share

Figure 8 displays the dynamics of the share of bank credit relative to total credit implied by the linear and non-linear model along with the data. All series are normalized to 1 at the

Table 5: Root Mean Square Error by horizons

(a) RMSE, Banks Net Worth

Forecast horizon	Linear	Non-Linear
4 quarter	0.285	0.168
7 quarter	0.298	0.221

(b) RMSE, Output

Horizon	Linear	Non-Linear
4 quarter	0.024	0.019
7 quarter	0.029	0.027

Note: Root mean square of forecasting errors from 2008:Q1 to 2015:Q1 for different horizons ($RMSE(h)$). $RMSE(h) = \sqrt{\sum_{T_0} \left(\hat{\epsilon}_{T_0}(T_0+h)\right)^2}$ with $\hat{\epsilon}_{T_0}(t) = log\left(\hat{obs}_{t/T_0}\right) - log\left(obs_t\right)$. $\hat{obs}_{t/T_0}$ is the predicted value of the observables at time t conditioning on the information set up to time T_0 delivered by system (25). Linear forecasts (blue line) sets $\phi_t = \bar{\phi}$ in system (25). Each model separately estimates the latent variables (endogenous states and exogenous shocks).

beginning of each recession. The loan share data comes from the U.S. Financial Accounts and corresponds to the ratio of non-financial corporate sector loans relative to total non-financial corporate sector debt securities and loans.

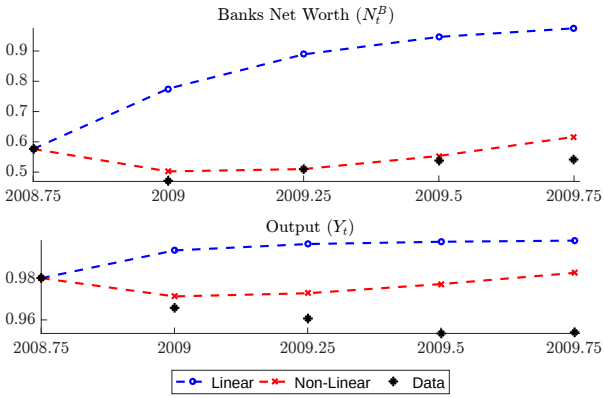
The data shows a considerable decline in the bank loans share during the crises of 1990 and 2008. For these particular periods the non-linear model is able to produce a decrease in the loan share of about 15% and 30% respectively, as opposed to the decrease of 1% and 5% generated by the linear model. In the model, the supply of credit is determined by banks' ability to intermediate. During financial crises, banks' net worth falls and reduces their capacity to bear aggregate risks. Banks' borrowing constraints tighten leading to a contraction in the supply of credit. Consequently, the bank loan share drops during financial crises. The intermediation mechanism exacerbates this effect, as the reduction of bank loans implies less diversification and collateral, leading to a further shrinkage of banks' balance sheets and credit supply.

Notice that the pattern generated by the nonlinear model is followed in the data with a certain lag. This lagged behavior might be explained by the difference in maturities on the different types of loans in the data, while in the model all contracts considered are short term. Another explanation might be that at the beginning of financial crises firms with pre-signed credit lines borrow up to their limits as a precautionary behavior to the expected contraction in credit, this is stressed for example in Becker and Ivashina [2014].

For the recession of 2001, when the intermediation mechanism is not present, both models display the same dynamics where the share of bank loans is unaffected. This pattern is followed in the data.

Figure 7: Forecast Analysis: Linear vs Non-linear dynamics

(a) Forecast from 2008



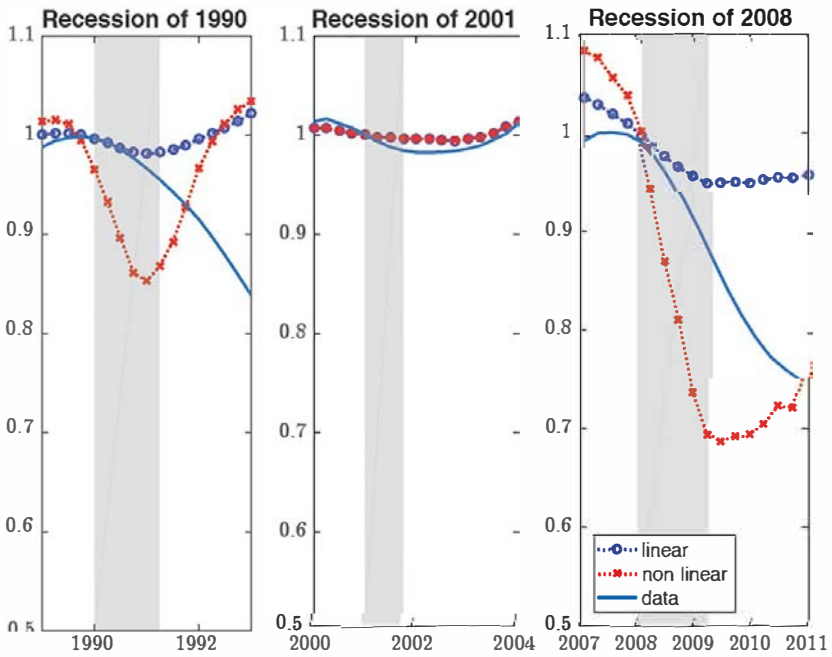
Note: Forecasts delivered by the system using information up to 2008:Q3 for the next four quarters, for banks' net worth and output, respectively, along with the realized observations (black diamonds). All series are shown as deviations from their linear trend. Linear forecasts (blue line) sets $\phi_t = \bar{\phi}$ in system (25). Each model separately estimates the latent variables (endogenous states and exogenous shocks).

7.3.2 The bank loan spread

Figure 9 plots the evolution of the cost of credit implied by the model non-linear dynamics and the observed evolution of two different measures of credit spreads (GZ spread and BAA-AAA) for the three recessions. The series are normalized to 1 the year prior to the beginning of each recession. As illustrated in Figure 9b, a distinct feature of the recession of 2008 is that the economy experienced a remarkable increase in the cost of credit. The crisis of 2001 also came along with a considerable increase in spreads but not as pronounced as the one of 2008. In contrast to the last two crises, the recession of 1990 did not present a substantial increase in the cost of credit.

Figure 9a shows the dynamics of the loan rate implied by the non-linear model in each of the recessions. The model produces similar patterns to the ones observed in the data, with a pronounced increase in 2008, a substantial but lower increase in 2001, and a moderated change in 1990. According to the model, the cost of credit delivers different paths depending on the type of recession. For instance, in non-financial recessions, firms' net worth and investment falls, and low levels of capital imply high marginal productivities and high returns on capital. Since banks share risks and returns with firms through loans, the loan rate increases as we see in the recession of 2001. In addition, during financial recessions, shocks to banks' balance sheets reduce their capacity to hold aggregate risks and contracts the supply of loans, implying a bigger increase in the loan rate, as the one displayed in the recession of 2008. The recession of 1990, even when it was a financial crisis, is associated with a smaller increase in loan rates because

Figure 8: Bank Loans Share during Recessions



Note: Share of bank loans in total firm debt implied by the model and data during the recessions of 1990, 2001, and 2008. The loan share data corresponds to the ratio of non-financial corporate business loans relative to total non-financial corporate sector debt securities and loans. Variables are scaled by their values at the beginning of each recession. All series are smoothed using a moving average filter (2 quarters). Linear dynamics (blue line) sets $\phi_t = \bar{\phi}$ in system (25) and re-estimates all latent variables.

the model implies a smaller drop in the level of capital and also a more negative technology shock, which induces a smaller change in returns and in the loan rate.

8 Conclusion

This paper provides a dynamic theory of financial intermediation where the distribution of net worth across different constrained borrowers (banks and firms) plays a role in real activity. In particular, firms are constrained borrowers with access to real investment opportunities, while banks are constrained borrowers that use their funds to finance firms. Both bank and firm net worth matter for real activity and can have a differential impact on investment, output, and interest rates. In particular, during financial recessions, the net worth of banks is special and has a stronger impact on investment and real activity.

The model assumes that banks are special because, due to their specific monitoring skills,

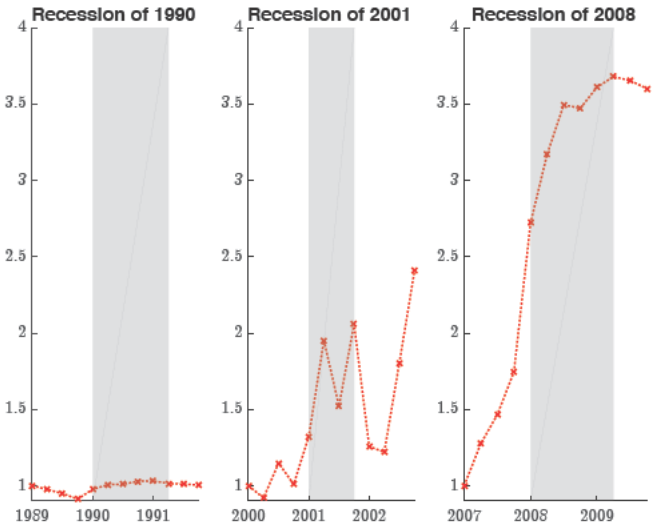
they are willing/able to share risks with firms that other lenders (households) are not. By lending to multiple firms, banks pool idiosyncratic risk and use their diversified portfolios as collateral to borrow from households. Bank intermediation allows additional funds to flow from households to firms. The net worth of firms determines their financing capacity, both from households and banks. The net worth of banks determines their debt capacity and, in turn, the amount of funds intermediated to firms.

During normal times, the net worth of both banks and firms affect real activity in the same way. This happens even when there are frictions between banks and firms. The insight is that, in this regime, banks can perfectly pass on to firms any increase in borrowing costs, generated by a drop in their net worth, by increasing their loan rates. Instead, during financial recessions, when the net worth of banks is critically low, loan rates are so high that banks cannot pass on to firms the increase in their borrowing costs. Therefore, a drop in bank net worth implies a stronger reduction in lending, which in turn impacts banking intermediation, by contracting banks' assets and the collateral provided to households. This further increases banks' borrowing costs and amplifies the initial shock. This positive feedback, generated by this intermediation mechanism, is a key innovation of the paper. The intermediation mechanism implies financial recessions are longer and more severe. A transfer of resources to banks or firms has a positive impact on investment, but recapitalizing banks is more effective. An increase in bank net worth has an associated multiplier effect that allows more funds to be channeled to firms, because of their diversification ability, than what an increase in firms net worth would induce.

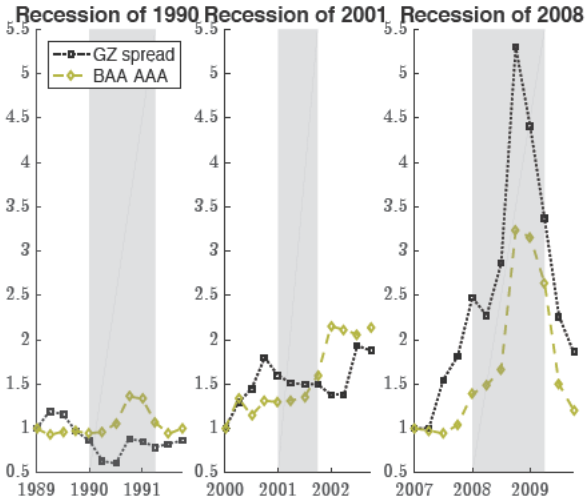
This new intermediation mechanism helps in explaining the observed non-linear relation between bank net worth and real economic activity. The estimation of the model finds that the intermediation mechanism is quantitatively important. In particular, this endogenous mechanism explains 40% of the fall in output and 80% of the fall in banks' net worth during the Great Recession. Moreover, it improves the forecasting performance of the model. Finally, the model generates dynamics that can explain the untargeted evolution of the share of bank loans in total firm debt and credit spreads during the recessions of 1990, 2001, and 2008.

Figure 9: Spread on Bank Loans during Recessions

(a) Model Dynamics: Loan Premium $E_t R^B_{t+1} - R$



(b) Data: Credit Spreads



Note: The spread on bank loans implied by the model and data for the recessions of 1990, 2001, and 2008. The model implied series corresponds to the loan premium which is the difference between the expected loan rate and the risk-free rate $E_t R^B_{t+1} - R$ estimated using the full model dynamics. The second panel shows two different measures of credit spreads: (i) the "GZ spread" corresponds to the average credit spread on senior unsecured bonds issued by non-financial firms constructed in Gilchrist and Zakrajšek [2011] ; (ii) the "BAA-AAA" corresponds to the spread between yields on Baa- and Aaa-rated long term industrial corporate bonds. Variables are scaled by their values a year prior to each recession. Shaded areas indicate NBER-dated recessions.

References

Tobias Adrian and Nina Boyarchenko. Intermediary leverage cycles and financial stability. *Becker Friedman Institute for Research in Economics Working Paper*, (2012-010), 2012.

Tobias Adrian, Emanuel Moench, and Hyun Song Shin. Financial intermediation, asset prices and macroeconomic dynamics. *FRB of New York Staff Report*, (422), 2010a.

Tobias Adrian, Emanuel Moench, and Hyun Song Shin. Macro risk premium and intermediary balance sheet quantities. *IMF Economic Review*, 58(1):179 207, 2010b.

Tobias Adrian, Erkko Etula, and Tyler Muir. Financial intermediaries and the cross-section of asset returns. *The Journal of Finance*, 69(6):2557 2596, 2014.

Bo Becker and Victoria Ivashina. Cyclicality of credit supply: Firm level evidence. *Journal of Monetary Economics*, 62:76 93, 2014.

Ben Bernanke and Mark Gertler. Agency costs, net worth, and business fluctuations. *The American Economic Review*, pages 14 31, 1989.

Ben S Bernanke and Alan S Blinder. The federal funds rate and the channels of monetary transmission. *The American Economic Review*, pages 901 921, 1992.

Ben S Bernanke and Mark Gertler. Inside the black box: The credit channel of monetary policy transmission. *The Journal of Economic Perspectives*, 9(4):27 48, 1995.

Ben S Bernanke, Mark Gertler, and Simon Gilchrist. The financial accelerator in a quantitative business cycle framework. *Handbook of macroeconomics*, 1:1341 1393, 1999.

Frédéric Boissay, Fabrice Collard, and Frank Smets. Booms and systemic banking crises. http://ssrn.com/abstract=2131075, January 2013. URL http://ssrn.com/abstract=2131075.

Michael Bordo, Barry Eichengreen, Daniela Klingebiel, and Maria Soledad Martinez-Peria. Is the crisis problem growing more severe? *Economic policy*, 16(32):52 82, 2001.

Markus Brunnermeier and Yuliy Sannikov. A macroeconomic model with a financial sector. *National Bank of Belgium Working Paper*, (236), 2012.

Markus K Brunnermeier and Yuliy Sannikov. A macroeconomic model with a financial sector. *The American Economic Review*, 104(2):379 421, 2014.

Markus K Brunnermeier, Thomas M Eisenbach, and Yuliy Sannikov. Macroeconomics with financial frictions: A survey. Technical report, National Bureau of Economic Research, 2012.

Guillermo A Calvo, Fabrizio Coricelli, and Pablo Ottonello. The labor market consequences of financial crises with or without inflation: Jobless and wageless recoveries. 2012.

Valerie Cerra and Sweta Chaman Saxena. Growth dynamics: the myth of economic recovery. *The American Economic Review*, 98(1):439 457, 2008.

Gabriel Chodorow-Reich. The employment effects of credit market disruptions: Firm-level evidence from the 2008 9 financial crisis. *The Quarterly Journal of Economics*, 129(1):1 59, 2013.

Lawrence J Christiano, Roberto Motto, and Massimo Rostagno. Risk shocks. *The American Economic Review*, 104(1):27 65, 2014.

Stijn Claessens, M Ayhan Kose, and Marco E Terrones. The global financial crisis: How similar? how different? how costly? *Journal of Asian Economics*, 21(3):247 264, 2010.

Sebastian Di Tella. Uncertainty shocks and balance sheet recessions. *Job Market Paper*, 2012.

Douglas W Diamond. Financial intermediation and delegated monitoring. *The Review of Economic Studies*, 51(3):393 414, 1984.

Douglas W Diamond and Raghuram G Rajan. A theory of bank capital. *The Journal of Finance*, 55(6):2431 2465, 2000.

Vadim Elenev, Tim Landvoigt, and Stijn Van Nieuwerburgh. A macroeconomic model with financially constrained producers and intermediaries. 2017.

Jesús Fernández-Villaverde and Juan F Rubio-Ramírez. Estimating macroeconomic models: A likelihood approach. *The Review of Economic Studies*, 74(4):1059 1087, 2007.

Mark Gertler and Peter Karadi. A model of unconventional monetary policy. *Journal of monetary Economics*, 58(1):17 34, 2011.

Mark Gertler and Nobuhiro Kiyotaki. Financial intermediation and credit policy in business cycle analysis. *Handbook of monetary economics*, 3(11):547 599, 2010.

Mark Gertler, Nobuhiro Kiyotaki, and Albert Queralto. Financial crises, bank risk exposure and government financial policy. *Journal of Monetary Economics*, 59:S17 S34, 2012.

Simon Gilchrist and Egon Zakrajšek. Credit spreads and business cycle fluctuations. Technical report, National Bureau of Economic Research, 2011.

Xavier Giroud and Holger M Mueller. Firm leverage and unemployment during the great recession. Technical report, National Bureau of Economic Research, 2015.

Luca Guerrieri and Matteo Iacoviello. Collateral constraints and macroeconomic asymmetries. *Journal of Monetary Economics*, 90:28 49, 2017.

Zhiguo He and Arvind Krishnamurthy. Intermediary asset pricing. *American Economic Review*, 103(2):732 70, 2013. doi: 10.1257/aer.103.2.732. URL http://www.aeaweb.org/articles.php?doi=10.1257/aer.103.2.732.

Zhiguo He and Arvind Krishnamurthy. A macroeconomic framework for quantifying systemic risk. Technical report, National Bureau of Economic Research, 2014.

Bengt Holmstrom and Jean Tirole. Financial intermediation, loanable funds, and the real sector. *the Quarterly Journal of economics*, pages 663 691, 1997.

R Glenn Hubbard. Capital-market imperfections and investment. *Journal of Economic Literature*, 36(1):193, 1998.

Matteo Iacoviello. Financial business cycles. *Review of Economic Dynamics*, 18(1):140 163, 2015.

Victoria Ivashina and David Scharfstein. Bank lending during the financial crisis of 2008. *Journal of Financial economics*, 97(3):319 338, 2010.

Urban Jermann and Vincenzo Quadrini. Macroeconomic effects of financial shocks. *The American Economic Review*, 102(1):238 271, 2012.

Òscar Jordà, Moritz Schularick, and Alan M Taylor. When credit bites back. *Journal of Money, Credit and Banking*, 45(s2):3 28, 2013.

Anil K Kashyap and Jeremy C Stein. Monetary policy and bank lending. In *Monetary policy*, pages 221 261. The University of Chicago Press, 1994.

Anil K Kashyap and Jeremy C Stein. What do a million observations on banks say about the transmission of monetary policy? *American Economic Review*, pages 407 428, 2000.

Nobuhiro Kiyotaki and John Moore. Credit cycles. *Journal of Political Economy*, 1997.

Stefan Krasa and Anne P Villamil. Monitoring the monitor: an incentive structure for a financial intermediary. *Journal of Economic Theory*, 57(1):197 221, 1992.

Arvind Krishnamurthy and Tyler Muir. How credit cycles across a financial crisis. Technical report, Stanford University Working Paper, 2016.

Zheng Liu, Pengfei Wang, and Tao Zha. Land-price dynamics and macroeconomic fluctuations. *Econometrica*, 81(3):1147 1184, 2013.

David López-Salido, Jeremy C Stein, and Egon Zakrajšek. Credit-market sentiment and the business cycle. Technical report, National Bureau of Economic Research, 2016.

Tyler Muir. Financial crises and risk premia. *Available at SSRN 2379608*, 2014.

Adriano A Rampini and S Viswanathan. Financial intermediary capital. Technical report, National Bureau of Economic Research, 2017.

Carmen M Reinhart and Kenneth S Rogoff. The aftermath of financial crises. Technical report, National Bureau of Economic Research, 2009.

Christina D Romer and David H Romer. New evidence on the impact of financial crises in advanced countries. Technical report, National Bureau of Economic Research, 2015.

Jeremy C Stein. Agency, information and corporate investment. *Handbook of the Economics of Finance*, 1:111 165, 2003.

Stephen D Williamson. Costly monitoring, financial intermediation, and equilibrium credit rationing. *Journal of Monetary Economics*, 18(2):159 179, 1986.

Appendix

A. Technology

Here I derive a case in which the return on assets of individual firms has constant returns to scale, while the aggregate production function has decreasing returns.

Consider a firm with a Cobb-Douglas production function which uses two factors of production k_i and h_i, we can think of capital and labor. Firms choose capital a period ahead, and hire labor at the same period of production but before the realization of the shock z_i. Lets denote the rental price at t of the factor $h_{i,t}$ with w_t (wage). However if the firm fails ($z_i = 0$), it defaults and doesn't pay wages. There is a competitive labor market where firms take wages as given. Thus, a firm that invested $k_{i,t-1}$ units of capital at $t-1$, maximizes cash-flows by choosing labor to solve

$$y_{i,t} = \max_{h_{i,t}} E_{t-1} z_{i,t} \left(k_{i,t-1}^{\alpha} h_{i,t}^{1-\alpha} - w_t h_{i,t} \right),$$

the optimal labor decision is linear on capital

$$h_{i,t}^*(k_{i,t-1}) = \left(\frac{1-\alpha}{w_t} \right)^{\frac{1}{\alpha}} k_{i,t-1}.$$

Thus, we can write the optimal cash-flows of the firm given a capital investment as

$$y_{i,t} = z_{i,t} A(w_t) k_{i,t-1},$$

with

$$A(w_t) = \alpha \left(\frac{1-\alpha}{w_t} \right)^{\frac{1-\alpha}{\alpha}}.$$

The aggregate demand of labor is

$$\int_i h_{i,t}^*(k_{i,t-1}) = \left(\frac{1-\alpha}{w_t} \right)^{\frac{1}{\alpha}} K_{t-1},$$

where $K_{t-1} = \int_i k_{i,t-1}$ denotes the aggregate capital.

The wage w_t clears the market of labor. Under an aggregate fixed labor supply which is normalized to one, $H_t = 1$, the equilibrium wage solves

$$1 = \left(\frac{1-\alpha}{w_t} \right)^{\frac{1}{\alpha}} K_{t-1},$$

which implies

$$A(w_t) = \alpha K_{t-1}^{\alpha-1}.$$

Therefore, while individual firms take wages as given and so their return on capital is linear, wages depend on aggregate capital and imply decreasing returns on aggregate. In particular, during recessions wages drop and so households are hit by the aggregate shocks. The drop in wage actually help firms as their profits per unit of capital increase. This is the specification used in the estimation in section 6.

B. Contracts

I restrict the analysis to short term contracts. A contract at t between the entrepreneur i, households and banks specifies how much each side should invest at t and how much it should be paid at $t+1$ as a function of the entrepreneur's report on $z_{i,t+1}$ which I denote with $\hat{z}$. No future or past reports can be used. The entrepreneur reports independently to households and banks.

Let's denote the total funds invested by the entrepreneur, households and banks with $N_{i,t}^E$, $B_{i,t}^H$ and $L_{i,t}^T$, respectively. Contracts must be incentive compatible. Thus, households must be promised a riskless return, otherwise the entrepreneur would always report $\hat{z} = 0$. Given banks monitoring skill, they can be promised a different return in case of success or failure, let's denote as $R_{t+1}^B(1)$ and $R_{t+1}^B(0)$ the returns promised to banks in case of success or failure, respectively. The following incentive compatibility condition enforces truthfull reporting

$$R_{t+1}^B(1)L_t^T - R_{t+1}^B(0)L_t^T \leq \bar{\phi}\kappa(1-\delta)k_{i,t},$$

where the left hand side is what the entrepreneur would get by misreporting failure in case of success and the right hand side the cost of misreporting. Moreover, the contract must satisfy that the promises are feasible, thus the following condition must hold

$$RB_{i,t}^H + R_{t+1}^B(0)L_{i,t}^T \leq (1-\kappa)(1-\delta)k_{i,t}.$$

Now, I show that any incentive compatible contract can be descentralized with riskless debt and a zero-recovery risky loan subject to the borrowing constraints described in section 6. First, note that we can replicate the promised payments to banks with a portofolio of a riskless bond with promise return $R_{t+1}^B(0)$ and a risky loan with promised $R_{t+1}^B(1) - R_{t+1}^B(0)$ in case of success and zero in case of failure. Then, the result follows by defining total riskless debt issued

$$B_{i,t} = \frac{RB_{i,t}^H + R_{t+1}^B(0)L_{i,t}^T}{R}$$

and zero-recovery loans

$$L_{i,t} = \frac{R^B_{t+1}(1)L^T_t - R^B_{t+1}(0)L^T_t}{R^B_{t+1}}.$$

The share of riskless debt held by households ($B^H_{i,t}$) or banks ($R^B_{t+1}(0)L^T$) is undetermined.

C. Omitted Proofs and Results

Proof of Proposition 1

Aggregate capital invested by firms follows by aggregating (9), which leads to $K_t = N^E_t + B_t + L_t$. Moreover, from (11) we have that $L_t = N^B_t + D_t$. Thus, total capital invested by firms is

$$K_t = N^E_t + N^B_t + B_t + D_t. \tag{26}$$

When both banks and firms borrow up to the maximum from households we have

$$B_t = \frac{(1-\kappa)(1-\delta)}{R} K_t \tag{27}$$

$$D_t = \frac{p_L R^B_{t+1} L_t}{R} \tag{28}$$

and when firms borrow up to the maximum from banks

$$R^B_{t+1} L_t = \bar{\phi}\kappa(1-\delta)K_t. \tag{29}$$

From (28) and (29) we get

$$D_t = \frac{p_L \bar{\phi}\kappa(1-\delta)}{R} K_t. \tag{30}$$

Finally, using (26), (27) and (30) we get

$$K_t = \frac{N^E_t + N^B_t}{1 - \frac{1}{R}\left(p_L\bar{\phi}\kappa + (1-\kappa)\right)(1-\delta)}.$$

By aggregating the laws of motion of firms and banks, (10) and (15), we get (20).

D. Estimation

Let $\mathbf{obs_t} = (Y_t, N_t^B)$ be the 2 x 1 vector of observables which includes the detrended time series of output and aggregate banks' net worth. Let $\mathbf{S_t} = (N_t, \beta_t, Z_t^Y, Z_t^{NB})$ be 4 x 1 vector of state variables, with $\mathbf{s_t} = (N_t, \beta_t)$ the 2 x 1 vector of endogenous state variables and $\mathbf{Z_t} = (Z_t^Y, Z_t^{NB})$ the 2 x 1 vector of exogenous states. Note that the state variables are latent. The state-space representation of the model is

$$\begin{aligned} \mathbf{obs_t} &= f(\mathbf{S_t}; \theta) + v_t \\ \mathbf{S_t} &= g\left(\mathbf{S_{t-1}}, \epsilon_\mathbf{t}; \theta\right). \end{aligned} \tag{31}$$

The functions f and g follow the system of equations in section 6.1. The first equation is the measurement equation, where v_t is a vector of Gaussian measurement errors. The second equation is the transition equation, which represents the law of motion for the state variables. The vector ϵ_t are the innovations to the exogenous shocks $\mathbf{Z}_t$. The likelihood function for the state-space model can be expressed as

$$\mathcal{L}(\mathbf{obs^T}; \theta) = \prod_{t=1}^{T} p(\mathbf{obs_t} | \mathbf{obs^{t-1}}; \theta)$$

where $\mathbf{obs^t} = [\mathbf{obs_1}, ..., \mathbf{obs_t}]$ and p denotes a density function. We can write

$$p(\mathbf{obs_t} | \mathbf{obs^{t-1}}; \theta) = \int p\left(\mathbf{obs_t} | \mathbf{S_t}; \theta\right) p\left(\mathbf{S_t} | \mathbf{obs^{t-1}}; \theta\right) d\mathbf{S_t}.$$

The conditional density of $\mathbf{obs_t}$ given $\mathbf{S_t}$ comes from the measurement equation and is Gaussian. I compute the probability of the latent state $\mathbf{S_t}$ given the information up to $t-1$ using a particle filter as in Fernández-Villaverde and Rubio-Ramírez [2007]. The recursive structure used to approximate the likelihood is summarized in the following pseudo-code:

Step 0, Initialization: Set $t = 1$. Initialize a vector of states $\left\{\mathbf{S}_0^i\right\}_{i=1}^N$ to their steady state values.

Step 1, Prediction: Sample N values of ϵ_t and compute $\left\{\mathbf{S}_{t|t-1}^i\right\}_{i=1}^N$ using $g(\mathbf{S}_{t-1}^i, \epsilon_t^i)$ from the transition equation in (31). This represents our approximated density $p\left(\mathbf{S_t} | \mathbf{obs^{t-1}}; \theta\right)$.

Step 2, Filtering: Assign to each draw $\mathbf{S}^i_{t|t-1}$ the particle weight

$$q^i_t = \frac{p\left(\mathbf{obs_t}|\mathbf{S^i_{t|t-1}};\theta\right)}{\sum_{i=1}^N p\left(\mathbf{obs_t}|\mathbf{S^i_{t|t-1}};\theta\right)}.$$

Step 3, Sampling: Sample N times from $\left\{\mathbf{S}^i_{t|t-1}\right\}_{i=1}^N$ with replacement and relative weights $\left\{q^i_t\right\}_{i=1}^N$. Call each draw $\mathbf{S}^i_t$. If $t < T$ set $t = t+1$ and go to step 1. Otherwise stop.

The likelihood function is then computed as

$$\mathcal{L}(\mathbf{obs^T};\theta) \approx \prod_{t=1}^T \frac{1}{N}\sum_{i=1}^N p\left(\mathbf{obs_t}|\mathbf{S^i_{t|t-1}};\theta\right).$$

The number of particles used in the estimation is $N = 150000$. The measurement errors v_t are included only for computational purposes and so are chosen to be small: the variances are equal to 0.0001 and the covariance is 0.

E. Summary of Model Mechanism

The model features three different kinds of agents: households, firms, and banks. Firms have access to productive projects. Firms can borrow from either households or banks, but have limited financing because their stochastic returns are privately observed, which gives rise to a moral hazard problem. Therefore, firms' net worth determines their financing capacity, both from households and banks. Households lend only through riskless debt. Banks are special because their monitoring skills allow them to make firms repay differently in good and bad states of the world, and thus extend risky loans (banks share risks). But, monitoring is private; thus households also only lend risk-free to banks. By lending to many firms, banks pool idiosyncratic risks and use their diversified portfolio of loans as collateral to borrow from households. This intermediation process allows funds to flow from households to firms through banks, funds that could not, otherwise, flow directly. However, banks hold aggregate risk on their balance sheets. This implies that they require net worth for intermediation, as they have to repay their debt even when a negative aggregate shock hits. As a result, the net worth of banks appears as an additional state variable, which can affect the economy differently than the net worth of firms.

Importantly, banks' monitoring skills are not perfect, there is a maximum repayment promise (in good states) that banks can enforce. This maximum promise is related to firms' collateral and so to the firms' net worth. Higher firm net worth reduces the incentives problems and increases this maximum promise, which allows firms to expand their bank credit. Hence, the net worth of firms determines their collateralizable assets and thus the amount they can borrow

from both households and banks. Banks net worth determines banks' intermediation capacity. Therefore, the net worth of both firms and banks affect investment: one through the demand of credit (firms' balance sheet channel) and the other through the supply of credit (lending channel).

The relative importance of each channel depends on the behavior of the loan rate. On the one hand, the payment promised to banks by firms consists of the amount of loans borrowed and the loan rate. A higher loan rate reduces the amount of loans firms can borrow and triggers a negative balance sheet channel. On the other hand, the constraint limiting the funds banks borrow from households depends on the return of banks' assets. Thus, a higher loan rate increases bank returns and positively affects the bank lending channel.

Moreover, the loan rate is endogenous. The loan rate clears the market for loans and so it is determined by the relative net worth of banks and firms. Thus, the relative sizes of net worth determine the importance of each of these channels in total credit and investment, and the economy features different dynamics depending on the level of the net worth of firms or banks.

Normal times are associated with the regions of the state space where firms borrow up to the limit of their constraints from both households and banks, and banks borrow up to the limit of their constraints from households. Financial frictions are always present in this regime. Nevertheless, the net worth of banks and firms play a similar role, and only their sum matters for investment and output. The standard financial accelerator appears, but this mechanism only depends on the total net worth of firms and banks.

In this region, a drop in banks' net worth triggers the lending channel: the drop tightens banks' constraints and raises the borrowing costs (shadow cost) for banks, contracting the loan supply. However, the scarcity of loans raises the loan rate which in turn helps banks and mitigates the effect. Such an increase in the loan rate affects firms' constraints and triggers the balance sheet channel affecting investment. It turns out that the effect on investment would be similar if firms had directly lost the net worth, instead of banks.

In this regime, we can think of a representative borrower that holds the total net worth of banks and firms and is constrained from households. For small shocks, the economy moves within this region. Therefore, we can represent banks and firms in one borrowing constrained sector, which has been the main focus of the literature.

However, the economy could also be in other regions in which the net worth of banks and firms have a differential effect. For instance, when the net worth of banks is critically low relative to firms, the loan rate that clears the market is too high relative to firms' expected return on capital, to the point where firms stop borrowing from banks up to the maximum of their constraints. This region describes a financial crisis.

During financial crises, the increase in banks' borrowing costs associated with a drop in banks' net worth cannot be passed on to firms anymore. The effects of the lending channel are not mitigated in this case because the loan rate is at the maximum at which firms can profitably borrow, and instead imply a more severe quantity adjustment in the amount of lending. Lower lending by banks implies a reduction in intermediation, as banks diversify over a lower amount of assets (loans), reducing the collateral used to borrow from households. Lower collateral tightens banks' constraints and implies an even higher increase in the borrowing costs for banks. This feedback loop, generated through the intermediation mechanism, leads to a collapse of banks' balance sheets and severely affects investment and output.

The key non-linear feature in the model is generated by the response of firms' loan demand to the corresponding contraction in the supply of loans. In normal times, firms demand is driven by the constraints. Firms are willing to borrow at higher loan rates and most of the adjustment from a drop in banks' net worth is made through prices. Instead, in financial crises, prices are already too high and most of the adjustment is made through quantities. A contraction in loan demand follows the initial contraction in loan supply.

Note that this *new intermediation mechanism* appears in addition to the standard financial accelerator: shocks that decrease banks' net worth, even if the net worth is redistributed to firms, imply a decline in intermediation and collateral, which prevents funds from flowing from households to firms and implies a reduction in investment and output. Moreover, the reduction of banks balance sheets reduces the exposure of banks to high assets returns. This induces additional persistence and implies a slow recovery of banks net worth and longer recessions.

Moreover, the model implies differences in the capital structure of firms and banks. Firms' debt is determined by the value of their assets that is not subject to idiosyncratic shocks. Banks, instead, diversify idiosyncratic shocks and their debt is determined by the lowest value of their assets which is only subject to aggregate shocks. The model implies that banks are more levered than firms, which makes them more sensitive to aggregate shocks. The ratio of banks net worth to firms net worth is pro-cyclical. This result implies that the economy moves to the financial crisis regime after a large enough negative shock to borrowers' assets.

Imprint and acknowledgements

I am grateful to Monika Piazzesi and Martin Schneider for their invaluable guidance. I also thank Saki Bigio, Sebastian Di Tella, Andres Drenik, Daniel Garcia-Macia, Eran Blass-Hoffman, Han Hong, Matteo Iacoviello, Pete Klenow, Arvind Krishnamurthy, Patrick Kehoe, Pablo Kurlat, Moritz Lenel, Claudio Michelacci, Diego Perez, Alessandra Peter, Cian Ruane, Anatoli Segura, Chris Tonetti and Lucciano Villacorta for helpful comments. Financial support from the Bradley Fellowship Program through a grant to SIEPR and the Macro Financial Modeling Initiative is acknowledged.

Alonso Villacorta
University of California, Santa Cruz, United States; email: **avillaco@ucsc.edu**

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website **www.esrb.europa.eu**

ISSN 2467-0677 (pdf)
ISBN 978-92-9472-020-7 (pdf)
DOI 10.2849/111676 (pdf)
EU catalogue No DT-AD-18-006-EN-N (pdf)